IG-GAN: Interactive Guided Generative Adversarial Networks for Multimodal Image Fusion

Chenhong Sui, *Member, IEEE,* Guobin Yang, Danfeng Hong, *Senior Member, IEEE,* Haipeng Wang, *Member, IEEE,* Jing Yao, *Member, IEEE,* Peter M Atkinson, Pedram Ghamisi, *Senior Member, IEEE*

***Abstract*—Multimodal image fusion has recently garnered increasing interest in the field of remote sensing. By leveraging the complementary information in different modalities, the fused results may be more favorable in characterizing objects of interest, thereby increasing the chance of a more comprehensive and accurate perception of the scene. Unfortunately, most existing fusion methods tend to extract modality-specific features independently without considering inter-modal alignment and complementarity, leading to a suboptimal fusion process. To address this issue, we propose a novel interactive guided generative adversarial network, named IG-GAN, for the task of multimodal image fusion. IG-GAN comprises guided dual streams tailored for enhanced learning of details and content, as well as cross-modal consistency. Specifically, a details-guided interactive running-in module and a content-guided interactive running-in module are developed, with the stronger modality serving as guidance for detail richness or content integrity, and the weaker one assisting. To fully integrate multi-granularity features from dual-modality, a hierarchical fusion and reconstruction branch is established. Specifically, a shallow interactive fusion module followed by a multi-level interactive fusion module is designed to aggregate multi-level local and long-range features. Concerning feature decoding and fused image generation, a high-level interactive fusion and reconstruction module is further developed. Additionally, to empower the fusion network to generate fused images with complete content, sharp edges, and high fidelity without supervision, a loss function facilitating the mutual game between the generator and two discriminators is also formulated. Comparative experiments with fourteen state-of-the-art methods are conducted on three datasets. Qualitative and quantitative results indicate that IG-GAN exhibits obvious superiority in terms of both visual effect and quantitative metrics. Moreover, experiments on two RGB-IR object detection datasets are also conducted, which demonstrate that IG-GAN can enhance the accuracy of object detection by integrating complementary information from different modalities.The code will be available at https://github.com/flower6top.**

***Index Terms*—Multimodal image fusion; Generative adversarial fusion; Transformer**

I. Introduction

Remarkable progress in sensor technology makes it possible to acquire multimodal images of the same scene [1]. However, influenced by sensor imaging mechanisms and the complex ground environment, single-mode images often cannot provide sufficient and detailed scene information [2]–[7]. For example, thermal infrared images contain the radiation signal from objects thus providing additional information to the visible spectrum but weak texture information [8], [9]. Synthetic aperture radar images (SAR) possess rich polarimetric scattering information of ground objects regardless of cloud cover but are seriously deficient in object details [10]. In comparison with infrared and SAR data, visible images are highly susceptible to weather and illumination variations despite providing great texture detail [8]. Therefore, multimodal image fusion is of great significance to simultaneously compensate for the content or detail deficiency of a single-mode sensor and enhance the information provided by images [11]. Fig. 1 presents an example of optical and SAR image fusion. As shown in Fig. 1, the SAR image can perceive objects in the red box, while its depiction of object details is weak. The optical image lacks complete perception of the objects and can only perceive the objects partially. For objects within the green box, the SAR image is unable to perceive objects that are subject to human interference, and the edges of the perceived objects are blurred, with unclear detailed features. In comparison, the optical image presents clearer object contours, and there is a significant difference in contrast between the object and the background.

Image fusion has benefitted from tremendous progress in the past decade [13]. Traditional fusion methods mainly depend on fixed transformations of source images and manually designed features for fusion [14]. These methods can be roughly divided into five categories(i.e., the multi-scale transform-based [15], the sparse representation-based [16], the subspace-based [17],

This work was supported by the National Key Research and Development Program of China under Grant 2022YFB3903401, the National Natural Science Foundation of China under Grant 42241109, 42271350, 61601397, 62076249, and the China Postdoctoral Science Foundation (Certificate Number: 2024M753307) (*Corresponding author: Danfeng Hong*)

C. Sui is with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China, also with the School of Physics and Electronic Information, Yantai University, Yantai 264005, China(e-mail: sui6662008@163.com).

G. Yang is with the School of Physics and Electronic Information, YanTai University, Yantai 264005, China, and also with the Weifang Vocational College, Weifang 261041, China (e-mail: yangguobing@s.ytu.edu.cn).

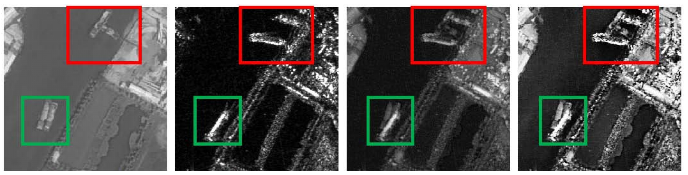
D. Hong is with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China, and also with the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100094, China (e-mail: hongdf@aircas.ac.cn).

H. Wang is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China, and also with the Institute of Information Fusion at Naval Aviation University, Yantai, China(e-mail: whp5691@163.com).

J. Yao is with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China. (e-mail: yaojing@aircas.ac.cn).

P. M. Atkinson is with the Faculty of Science and Technology, Lancaster University, Lancaster, U.K. (e-mail:pma@lancaster.ac.uk).

P. Ghamisi is with the Helmholtz-Zentrum Dresden-Rossendorf (HZDR), 09599 Freiberg, Germany, and he is also with Lancaster University, Lancaster, U.K. (e-mail: p.ghamisi@gmail.com).



(a) Optical (b) SAR (c) GANMcC (d) IG-GAN

Fig. 1. An Example of optical and SAR image fusion. (**a**) Optical Image; (**b**) SAR image; (**c**) Fusion result of GANMcC [12]; (**d**) Fusion result from IG-GAN. It is evident that compared to GANMAC, IG-GAN can leverage the advantages of SAR and optical images to achieve a more complete perception of objects within red and green boxes, and fused images have significant advantages in sharpening edges and contrast.

the saliency-based [18], and the hybrid method [19]). Despite the high efficiency of most traditional methods, artificially designed feature extraction or fusion rules cannot fully capture the characteristics of different modalities, resulting in limited fusion performance.

Deep fusion methods, as a new paradigm beyond traditional methods, rely on the powerful modeling capabilities of deep neural networks for adaptive information extraction and learning the fusion rules [20]. Examples include convolutional neural networks (CNN) [21] and autoencoders (AE) [22], [23] for modality-specific feature extraction and fusion. For example, [24] presents the CNN-based pansharpening methods with three convolutional layers, which can integrate the unique advantages of a panchromatic (PAN) image and a multispectral (MS) image. [25] introduces the dual-attention to inject the spatial details of PAN into a hyperspectral (HS) image for superresolution. [26] adopts the encoder-decoder-based architecture for HS and Lidar fusion. U2Fusion [27] proposes an unsupervised fusion method based on a densely connected network for various image fusion tasks. RFN-Nest [28] introduces an end-to-end residual architecture-based fusion network for infrared and visible image fusion. For the complementary fusion of texture and intensity information, PMGI [29] advocates constructing a dual-path-based unified fusion network for better information extraction. Additionally, since a generative adversarial network (GAN) is capable of generating high-fidelity fusion images through the min-max game between the generator and discriminator, it is naturally utilized for multi-modal image fusion in various studies [4], [30]–[32]. For example, [33] develops a GAN-based two-stage framework for spatiotemporal image fusion. [34] establishes a GAN-based network for SAR-Optical fusion and cloud removal. Additionally, in [35], GAN is successfully applied to HS pansharpening, which adopts the 3D convolutional network to capture desirable high-frequency residuals. FusionGAN [36] establishes an adversarial game for the fusion of infrared and visible images, effectively avoiding the artificial design of fusion rules. DDcGAN [37] extends FusionGAN by adding a discriminator, which helps the fused image fully retain information from multi-source images. GANMCc [12] recognizes that the details and contrast in the infrared image are not as significant as in the visible image. Therefore, it employs multiple classifiers as discriminators to output the probability that the input is an infrared image or a visible image, producing a visually appealing fused image. These deep fusion methods have significantly advanced the field of image fusion. Nevertheless, they are limited by the inherent constraints of convolutional operators in capturing inter- or intra-modality global context.

To address the above issue, some researchers have turned to the Transformer structure [38], which embraces long-term modeling capabilities. For example, IFT [39] introduces the Transformer to image fusion and achieves performance similar to the CNN architecture. [40] introduces the Transformer to HS-MS fusion, where the structured embedding matrix is injected into the Transformer encoder to learn the residual map. [41] leverages the Transformer for HS pansharpening, in which the modality-specific feature extractors are designed to capture textural details. [42] proposes an attention-based multiscale transformer network to model contextual information in bi-temporal images for change detection. [43] proposes the center attention transformer (CAT) with a stratified spatial-spectral token for HSI classification. [44] learned a coupling model-driven and data-driven paradigm to distinguish between the background and anomalies for hyperspectral anomaly detection. TGFuse [45] embeds the Transformer into GAN for global visible and infrared image fusion. SwinFusion [46] utilizes the Swin Transformer [47] to extract features from different sources and leverages cross-domain attention for feature fusion.

For the Transformer-based methods, the construction of a global feature association is beneficial leading to better-fused images [48]. However, the approach generally conducts semantic mining for each modal independently, with no perception of the inter-modal discrepancy and consistency. This could result in the underutilization of the modality-specific advantages and cross-modality commonalities while the overusing of invalid or noisy information.

To the above end, this paper introduces an Interactive Guided Generative Adversarial Fusion Network for multi-modal images (IG-GAN). Note that some modalities are capable of collecting complete scene information irrespective of weather or illumination variations, whereas visible images possess the merit of rich texture details. Enlightened by this, detail and content streams are first cooperatively established rather than independently for cross-modal complementarity and consistency enhancement. Specifically, a details-guided interactive running-in module (GIR$_1$) and a content-guided interactive running-in module (GIR$_2$) are developed. This is conducive to ensuring that the advantages of the dominant modality can be fully utilized, while the other modality can assist in cross-modal feature alignment, enhancement, and complementary fusion. Regarding the comprehensive integration of dual-stream features, we further construct a hierarchical fusion and reconstruction branch. In this branch, both a shallow interactive fusion module (SIF) and a multi-level interactive fusion module (MIF) are built. Furthermore, for fine decoding and fused image generation, we propose a high-level interactive fusion and reconstruction module capable of absorbing multi-modal, multi-granularity local-global features. Additionally, to guarantee that the fusion network can generate complete, sharpened, and high-fidelity images, we design a loss function involving the mutual game between the generator and two discriminators. Qualitative and quantitative experimental results show that compared with state-of-the-art methods, IG-GAN exhibits apparent superiority over others on four commonly used benchmarks. Our main contributions can be summarized as follows.

- A novel unsupervised multimodal image fusion method, called IG-GAN, is proposed to fully explore the modal-specific advantageous information regarding detail richness and content completeness while enhancing cross-modal commonalities collaboratively.
- In the generator, a guided details stream and a guided content stream are established for multi-level inter-modality alignment, cooperation, and enhancement. Specifically, a details-guided interactive running-in module (GIR$_1$) and a content-guided interactive running-in module (GIR$_2$) are developed. This means that both the content and detail streams are built with multi-modal interactive promotion rather than operating independently. In each stream, both the leading role of each dominant modality and the auxiliary contribution of the weaker one are considered.
- Concerning dual-stream feature integration, the generator also involves a shallow interactive fusion module (SIF), a multi-level interactive fusion module (MIF), and a high-level fusion and reconstruction module (HRM). This promotes the multi-granularity, multi-level integration of dual-stream features, and fused image generation.
- To ensure the fusion performance of our IG-GAN without supervision, a novel loss function simultaneously involving detail richness, content integrity, and high fidelity is devised for network training. It boosts the mutual game between the generator and two discriminators. Qualitative and quantitative experimental results on four benchmark datasets demonstrate the superiority of IG-GAN.

The remainder of the paper is organized as follows. Section II primarily reviews typical work related to IG-GAN. Section III gives a detailed description of IG-GAN and its core modules(e.g., GIR$_1$, GIR$_2$, SIF, MIF, and HRM). In Section IV, quantitative and qualitative experimental results and discussion are provided. Ultimately, the conclusions and future research are presented in Section V.

II. Related Works

A. GANMcC

In the context of infrared and visible image fusion, it is commonly acknowledged that the former lacks visual information such as details and textures, but can effectively depict significant objects under low illumination. Conversely, the latter is rich in detailed texture but is highly susceptible to changes in lighting conditions. Under low illumination, scene information in visible images may appear incomplete or even completely missing with low contrast. To address the integration of saliency and details from infrared and visible images, a multi-classification generative adversarial network (GANMcC) [12] is designed.

The main idea of GANMcC is to transform multi-modal image fusion into a simultaneous estimation of the contribution from infrared and visible image distributions. GANMcC comprises a generator and two discriminators. The generator aims to maximize the probability that the fused image comes from both visible and infrared images. In contrast, the discriminator adopts a multi-class classifier to determine that the fused image is neither an infrared image nor a visible image. With continuous adversarial learning, the generator can estimate the probability distribution of both infrared and visible images, enabling the generation of fused images with significant contrast and rich texture details. Eq. (1) depicts the loss function in favor of both the generator and discriminator.

$$\mathcal{L}{=}\mathcal{L}_G{+}\mathcal{L}_D \tag{1}$$

where $\mathcal{L}_G$ is the loss of the generator, $\mathcal{L}_D$ corresponds to the loss of the discriminators.

$\mathcal{L}_G$ consists of both content loss $\mathcal{L}_{G_{con}}$ and adversarial loss $\mathcal{L}_{G_{adv}}$. Regarding $\mathcal{L}_{G_{con}}$, to preserve the details and textures in multi-modal images, intensity loss, and gradient loss are provided as shown in Eq. (2).

$$\begin{aligned}\mathcal{L}_{G_{\text{con}}} = \beta_1 \left\|\boldsymbol{I_f} - \boldsymbol{I}_{ir}\right\|_F^2 + \beta_2 \left\|\nabla\boldsymbol{I_f} - \nabla\boldsymbol{I}_{vis}\right\|_F^2 \\ + \beta_3 \left\|\nabla\boldsymbol{I_f} - \nabla\boldsymbol{I}_{ir}\right\|_F^2 + \beta_4 \left\|\boldsymbol{I_f} - \boldsymbol{I}_{vis}\right\|_F^2\end{aligned} \tag{2}$$

where $\boldsymbol{I_f}$, $\boldsymbol{I}_{ir}$, and $\boldsymbol{I}_{vis}$ represent the fused image, the infrared image, and the visible image, respectively. ∇ denotes the second-order gradient operator. $\|\cdot\|_F$ is the $Frobenius-norm$. β_1, β_2, β_3, and β_4 are the trade-off parameters.

The adversarial loss $\mathcal{L}_{G_{adv}}$ can be described as:

$$\mathcal{L}_{G_{adv}} = (D(\boldsymbol{I_f})[1] - d)^2 + (D(\boldsymbol{I_f})[2] - d)^2 \tag{3}$$

where D is the discriminator, in which d is the image modal label used to determine the type of fused image. $D(\cdot)[1]$ and $D(\cdot)[2]$ represent the probability that the fused image is the visible image or the infrared image, respectively.

Concerning discriminators, they are capable of judging that the distribution of the fused image is different from both the visible and infrared images. Therefore, the corresponding discriminator loss $\mathcal{L}_D$ is defined as:

$$\mathcal{L}_D = \mathcal{L}_{D_{\text{vis}}} + \mathcal{L}_{D_{\text{ir}}} + \mathcal{L}_{D_{\text{fused}}} \tag{4}$$

where $\mathcal{L}_{D_{\text{vis}}}$ aims to measure the probability that the visible image is classified as an infrared image. Analogously, $\mathcal{L}_{D_{\text{ir}}}$ depicts the probability that the infrared image is classified as a visible image. Meanwhile, $\mathcal{L}_{D_{\text{fused}}}$ represents the probability that the fused image is classified as a visible or infrared image.

The mutual game between the generator and the discriminators, based on a loss function, is beneficial for producing a high-quality fused image. Unfortunately, due to the local modeling attributes of convolutional operators, the exploration of inter- or intra-modality global semantics is lacking.

B. SwinFusion

SwinFusion [46] is a versatile image fusion framework based on cross-domain distance learning and the Swin Transformer [47]. Unlike existing transformer-based image fusion methods that mainly focus on the interaction of information within a domain, SwinFusion takes a step further by exploring the contextual relationship between multi-source images. To address this limitation, SwinFusion employs the Swin Transformer as the backbone to model long-range dependencies between domains and design cross-domain attention for feature fusion. Specifically, the Swin Transformer is introduced to delve deeper into the semantics extracted by a CNN from shallow features. Subsequently, cross-attention is utilized for effective feature fusion. This approach ensures the integration of cross-domain context information, leading to exceptional image fusion results. The total loss, denoted as $\mathcal{L}_{total}$, is composed of SSIM loss $\mathcal{L}_{ssim}$ [49], texture loss $\mathcal{L}_{tex}$, and intensity loss $\mathcal{L}_{int}$ as described in Eq. (5).

$$\mathcal{L}_{total} = \mathcal{L}_{ssim} + \mathcal{L}_{text} + \mathcal{L}_{\mathrm{int}} \tag{5}$$

where the SSIM loss $\mathcal{L}_{ssim}$ can be expressed as

$$\mathcal{L}_{ssim} = w_1(1 - ssim(\boldsymbol{I_f}, \boldsymbol{I_1})) + w_2(1 - ssim(\boldsymbol{I_f}, \boldsymbol{I_2})) \tag{6}$$

where $ssim(\cdot)$ represents the structural similarity operation, the balancing parameters w_1 and w_2 are set to 0.5. $\boldsymbol{I_1}$ and $\boldsymbol{I_2}$ denote the bi-modal images, and $\boldsymbol{I_f}$ is the fused image.

To characterize the richness of texture in images, Eq. (7) further provides the texture loss $\mathcal{L}_{tex}$.

$$\mathcal{L}_{tex} = \frac{1}{HW}\left\| |\nabla \boldsymbol{I_f}| - \max(|\nabla \boldsymbol{I_1}|, |\nabla \boldsymbol{I_2}|) \right\|_1 \tag{7}$$

where $|\cdot|$ denotes the absolute operation, $\|\cdot\|_1$ is $l_1 - norm$, and $\max(\cdot)$ refers to the element-wise maximum selection. H and W denote the height and width of the image.

Additionally, SwinFusion adopts the intensity loss L_{int} to describe the element-by-element spatial discrepancy between the fusion image and the original bimodal images, as expressed in Eq. (8)

$$\mathcal{L}_{\mathrm{int}} = \frac{1}{HW}\left\| \boldsymbol{I_f} - M(\boldsymbol{I_1}, \boldsymbol{I_2}) \right\|_1 \tag{8}$$

where $M(\cdot)$ is an element-wise aggregation operation.

Note that the SwinFusion is an encoder-decoder-based fusion network, which adopts the SSIM loss, text loss, and intensity loss to optimize the fusion network. In comparison, many GAN-based fusion methods [12], [50] leverage both the generator loss and the discriminator loss for fusion network optimization, which contribute to generating high-quality fused images, ensuring the naturalness and realism of the fusion results.

When delving deeper into the SwinFusion model, we noticed that the model adopts a cross-attention mechanism for feature fusion in the second stage of the Swin Transformer, but lacks independent fusion branches. Consequently, this may limit the flexibility and efficiency of the model when dealing with complex data. In this regard, it is sensible to introduce an independent fusion branch to enhance the overall performance and adaptability of the model.

III. Our Method

In this section, we first introduce the framework of our proposed IG-GAN. Then, the specific modules in the generator and discriminator networks are described, respectively. Finally, to optimize the designed network and enable it to produce complete and detailed images without supervision, a corresponding loss function is provided.

A. Framework of IG-GAN

To yield high-quality fused images, multimodal image fusion needs to explore and utilize inter-modal semantic consistency for better feature alignment and enhancement. Additionally, inter-modality complementary information is required to compensate for the deficiencies of single-source images. To address this, we propose a dual-stream interactive guided generative adversarial fusion network (IG-GAN) through the mutual game between the generator and discriminators, as depicted in Fig. 2.

As depicted in Fig. 2, the generator comprises a guided details stream, a guided content stream, and a hierarchical feature fusion and image reconstruction branch. The former primarily focuses on complementary feature mining and aligned feature enhancement, while the latter is mainly responsible for further multi-view and multi-level feature fusion, as well as fused image restoration. Specifically, in each stream, the guided interactive running-in modules (i.e., GIR_1 and GIR_2) are introduced based on four consecutive guided Swin Transformer blocks. Concerning hierarchical fusion, we first establish a shallow interactive fusion module (SIF) for the generation of multi-view and low-level fusion features $\boldsymbol{S_f}$. Subsequently, a multi-level interactive fusion module (MIF) is utilized to produce both low and high-level fusion features $\boldsymbol{M_f}$. Following the ResNet approach, we advocate feeding both low-level and high-level fusion features into an aggregation block that involves concatenation and a 1×1 convolution. This is followed by three consecutive Transformer blocks. Finally, after the patch expansion operation in the Transformer, the aggregation block, accompanied by multiple Transformer modules, serves as the fusion feature decoding and fused image reconstruction.

As an unsupervised image fusion network, two discriminators, namely $\mathrm{Discriminator}_1$ and $\mathrm{Discriminator}_2$, are employed to distinguish between the fused images and source images. They play a crucial role in preventing significant discrepancies between the fused and original images and avoiding artifacts. Therefore, in Section III-B, we first introduce the guided dual-stream and hierarchical fusion and image restoration parts in the generator, respectively.

Influenced by the working principle of the sensor, there are significant differences in the details (such as texture and contrast) and integrity of various modal images in bad weather and low illumination. For example, SAR images can capture complete scenes unaffected by clouds, rain, fog, and lighting changes. Unfortunately, they lack texture and contrast information. On the other hand, optical images are typically rich in detailed textures, but are susceptible to weather and illumination changes, leading to potential severe loss or pollution of scene information. Therefore, it is wise to explore and aggregate their complementary semantics in terms of rich details and content integrity, respectively. In this context, we construct a dual-stream architecture for better feature extraction and fusion,

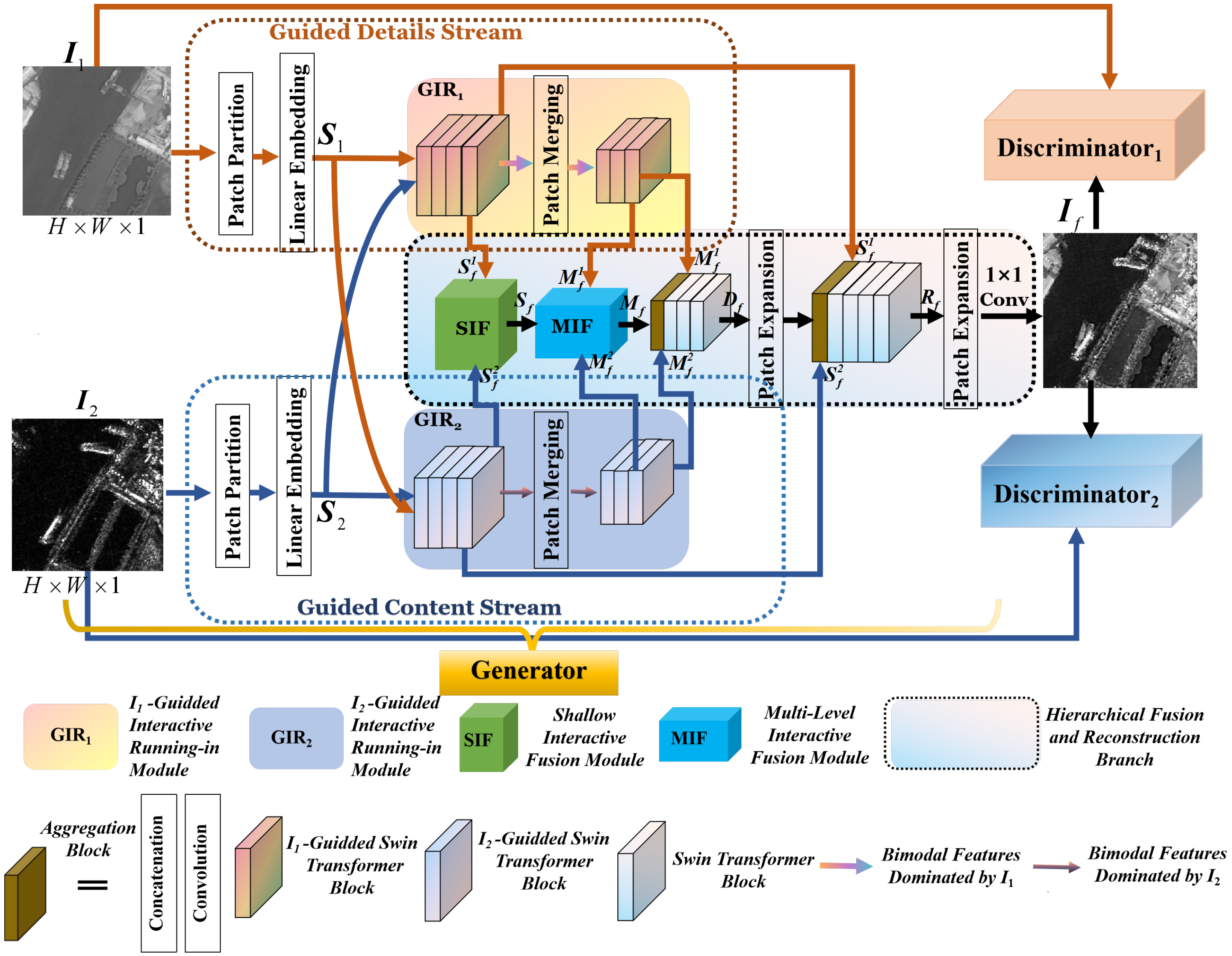


Fig. 2. Framework of IG-GAN. As a generative fusion network, IG-GAN has $\boldsymbol{I_1}$-guided details stream, $\boldsymbol{I_2}$-guided content stream, as well as multimodal hierarchical fusion and a reconstruction branch in the generator. This facilitates the comprehensive exploration, alignment, and enhancement of cross-modal semantic consistency. Meanwhile, modality-wise different advantages respecting detail richness and content completeness are fully considered. To enhance the fidelity of the fused images, dual discriminators are leveraged to engage in a game against the generator.

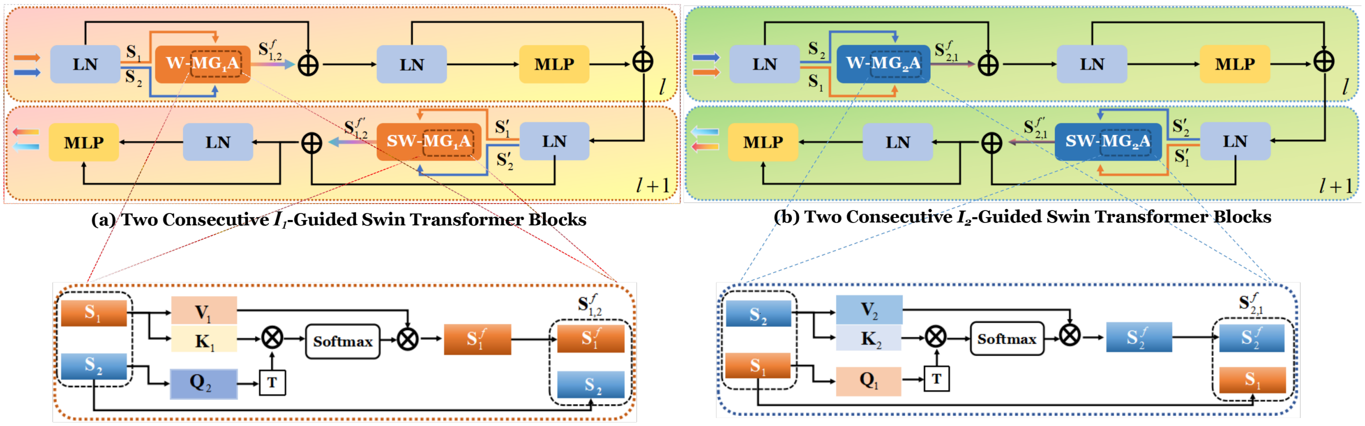
where different modalities dominate the details stream and content stream based on their performance advantages.

Let $\boldsymbol{I_1}$ be the input optical image with size $H \times W \times 1$, while $\boldsymbol{I_2}$ denotes the corresponding infrared or SAR image with the same size. Then, as shown in Fig. 2, $\boldsymbol{I_1}$ should play a dominant role in the details stream, as it is beneficial for retaining contextual details. Regarding $\boldsymbol{I_2}$, aligning spatial and semantic information with $\boldsymbol{I_1}$ allows $\boldsymbol{I_1}$ to assist $\boldsymbol{I_2}$ in capturing more complete scene content while preserving effective detail information. For the content stream, $\boldsymbol{I_2}$ should assume a critical role, providing more comprehensive scene content that remains unaffected by lighting and weather conditions. In this context, $\boldsymbol{I_1}$ has the auxiliary effect of enhancing detailed textures by aligning spatially and semantically with $\boldsymbol{I_2}$.

B. Guided Dual-Stream

Note that inter-modality features have both uniqueness and relevance, involving some common semantic information. In this view, compared with common independent feature extraction mechanisms from various modalities, it is sensible to explore cross-modal information cooperatively, which helps strengthen the heterogeneous or complementary information. In this regard, we give a $\boldsymbol{I_1}$-guided interactive running-in module (GIR$_1$) and a $\boldsymbol{I_2}$-guided interactive running-in module (GIR$_2$) for comprehensive feature alignment, cooperation, and consistency feature enhancement.

1) Guided Interactive Running-in Module: The guided interactive running-in module aims to explore the commonality and uniqueness among modalities and allows multimodal images to be included jointly. As described in Fig. 2, after passing through the U-shaped network, the features from $\boldsymbol{I_1}$ and $\boldsymbol{I_2}$ are concatenated to form joint features in two streams.



(a) Two Consecutive I_1-Guided Swin Transformer Blocks

(b) Two Consecutive I_2-Guided Swin Transformer Blocks

Fig. 3. Illustration of the proposed $\boldsymbol{I_1}$- and $\boldsymbol{I_2}$-Guided Swin Transformer blocks used in GIR$_1$ and GIR$_2$, respectively. Different from the standard Swin Transformer block only involving a single modality as input, the guided ones corresponding to (a) and (b) take the features from dual-modality as input. In specific, for (a), owing to $\boldsymbol{I_1}$-guided multi-head attention in W-MG$_1$A and SW-MG$_1$A, the outputs of details stream are mainly dominated by features from $\boldsymbol{I_1}$, whereas the ones from $\boldsymbol{I_2}$ helps feature correction and enhancement via cross-modal feature interaction alignment. Regarding (b), $\boldsymbol{I_2}$-guided multi-head attention is given in W-MG$_2$A and SW-MG$_2$A. In this case, $\boldsymbol{I_2}$ dominates the content stream, while $\boldsymbol{I_1}$ helps to strengthen consistency features through feature interaction alignment.

Then, they are fed into GIR$_1$ and GIR$_2$ with the first half of the channels dominant and the second half of the channels auxiliary. This ensures that $\boldsymbol{I_1}$ and $\boldsymbol{I_2}$ play different roles in the two streams based on their contribution. Compared with independent feature extraction from each modality, this is conducive to aligning and enhancing the commonality between modalities. Moreover, heterogeneous or complementary information is strengthened.

Note that the Swin Transformer possesses multi-view perception and long-range modeling ability [51]. Additionally, with the utilization of patch merging, consecutive Swin Transformer modules can capture semantic features at lower scales. Motivated by the merits of the Swin Transformer, we have developed the $\boldsymbol{I_1}$-guided and $\boldsymbol{I_2}$-guided Swin Transformer blocks used in GIR$_1$ and GIR$_2$ for full interaction alignment and enhancement for $\boldsymbol{I_1}$ and $\boldsymbol{I_2}$. Specifically, inspired by cross-attention, Swin Transformer's multi-head attention is improved to $\boldsymbol{I_1}$-guided and $\boldsymbol{I_2}$-guided multi-head attention mechanisms.

As demonstrated in Fig. 2, for the guided details stream, $4 \times$ $\boldsymbol{I_1}$-guided Swin Transformer blocks are built to extract detailed semantic information dominated by optical images. Meanwhile, the features from infrared or SAR images are leveraged to assist in semantic alignment, correction, and consistency enhancement. Note that patch merging is an effective resolution reduction mechanism that does not cause information loss [39]. Therefore, to capture global semantics at larger scales, the patch merging operation is leveraged in both GIR$_1$ and GIR$_2$. Assuming that $\boldsymbol{S_1}$ and $\boldsymbol{S_2}$ have a spatial size of $\frac{H}{4} \times \frac{W}{4}$ after patch partition, then, after patch merging, the spatial size of the feature map is reduced to $\frac{H}{8} \times \frac{W}{8}$. After that, $3 \times \boldsymbol{I_1}$-guided Swin Transformer blocks are further adopted to extract semantic information.

Analogous to the details stream, for the content stream, a backbone content extraction module is constructed, which contains $4\times$ $\boldsymbol{I_2}$-guided Swin Transformer blocks dominated by SAR or infrared images. At the same time, the features from optical images are also introduced into this stream, which play an auxiliary role in semantic consistency enhancement via interactive alignment. After a patch merging operation, $3\times$ $\boldsymbol{I_2}$-guided Swin Transformer blocks are capable of exploring interaction features with size $\frac{H}{8} \times \frac{W}{8}$.

For ease of understanding, before analyzing the proposed guided attention mechanisms and Swin Transformer, we first recall the standard ones. Let $\boldsymbol{S}$ be the input. Then, the multi-head self-attention $MHead(\boldsymbol{S})$ in W-MSA and SW-MSA of the standard Swin Transformer blocks can be described as

$$MHead(\boldsymbol{S}) = Concat(\boldsymbol{head}^1, \boldsymbol{head}^2, \cdots, \boldsymbol{head}^h)\boldsymbol{W^O} \tag{9}$$

where $\boldsymbol{head}^i$ represents the i^{th} head of $MHead(\boldsymbol{S})$. h denotes the number of heads. $\boldsymbol{head}^i$ can be obtained as defined in Eq. (10).

$$\boldsymbol{head}^i = soft\max\left(\frac{\boldsymbol{K}^i(\boldsymbol{Q}^i)^T}{\sqrt{d_k}} + \boldsymbol{B}\right)\boldsymbol{V}^i \tag{10}$$

where $\boldsymbol{Q}^i$,$\boldsymbol{K}^i$,$\boldsymbol{V}^i$ correspond to different linear mappings of $\boldsymbol{S}$, which satisfy $\boldsymbol{Q}^i = \boldsymbol{S} \cdot \boldsymbol{W_i^Q}$, $\boldsymbol{K}^i = \boldsymbol{S} \cdot \boldsymbol{W}_i^K$, $\boldsymbol{V}^i = \boldsymbol{S} \cdot \boldsymbol{W}_i^V$. d_k denotes the dimension of $\boldsymbol{K}^i$, $\boldsymbol{B}$ is the relative position encoding that can be learned.

As shown in Eq. (10), despite the global statistics of the standard Swin Transformer, the existing multi-head self-attention tends to focus on a single source input $\boldsymbol{S}$ [46], [52]. Consequently, the association between modalities is neglected, which is not conducive to the enhancement of inter-modality consistency and the fusion of complementarity. In this connection, it is necessary to make reasonable use of the auxiliary role of other modalities while affirming the dominant role of stronger modalities. This means that we should start with the importance of each mode, and delegate our attention to the important leading mode, while other modes assist in enhancing.

For better illustration, Fig. 3 provides a visual representation of the $\boldsymbol{I_1}$- and $\boldsymbol{I_2}$-guided Swin Transformer blocks.

Specifically, the $\boldsymbol{I_1}$- and $\boldsymbol{I_2}$-guided attention mechanisms are displayed as well.

As described in Fig. 3, the guided Swin Transformer blocks contribute to the mutual running-in of cross-modality features via a bimodal guided attention mechanism. For example, W-MG$_1$A and SW-MG$_1$A contain the $\boldsymbol{I_1}$-guided multi-head attention module. In this module, the features of $\boldsymbol{I_1}$ play a crucial role, whereas those of $\boldsymbol{I_2}$ assist in aligning cross-modal features and enhancing consistency features. In Fig. 3 (b), W-MG$_2$A and SW-MG$_2$A involve the $\boldsymbol{I_2}$-guided multi-head attention module. Here, the features from $\boldsymbol{I_2}$ dominate, while those from $\boldsymbol{I_2}$ are used for cross-modal feature alignment and enhancement. The $\boldsymbol{I_1}$-guided multi-head attention in W-MG$_1$A and SW-MG$_1$A can be described as follows.

$$\begin{aligned} G_1MHead(\boldsymbol{S_1}, \boldsymbol{S_2}) = Concat(&\boldsymbol{g_1head}^1, \boldsymbol{g_1head}^2, \dots, \\ &\boldsymbol{g_1head}^h) \cdot \boldsymbol{W^O} \end{aligned} \tag{11}$$

where $\boldsymbol{g_1head}^i$ denotes the i^{th} $\boldsymbol{I_1}$-guided head in $G_1MHead(\boldsymbol{S_1}, \boldsymbol{S_2})$, $Concat$ means the concatenation operation, $\boldsymbol{W^O}$ is a linear embedding matrix.

Analogously, the $\boldsymbol{I_2}$-guided multi-head attention $G_2MHead(\boldsymbol{S_2}, \boldsymbol{S_1})$ in W-MG$_2$A and SW-MG$_2$A can be depicted as

$$\begin{aligned} G_2MHead(\boldsymbol{S_2}, \boldsymbol{S_1}) = Concat(&\boldsymbol{g_2head}^1, \boldsymbol{g_2head}^2, \dots, \\ &\boldsymbol{g_2head}^h) \cdot \boldsymbol{W^O} \end{aligned} \tag{12}$$

where $\boldsymbol{g_2head}^i$ denotes the i^{th} $\boldsymbol{I_2}$-guided head in $G_2MHead(\boldsymbol{S_2}, \boldsymbol{S_1})$.

Inspired by the cross-attention, we give the specific formulas of guided attention as follows.

$$\begin{cases} \boldsymbol{g_1head}^i = \text{softmax}\left(\dfrac{\boldsymbol{K}_1^i \cdot (\boldsymbol{Q}_2^i)^\intercal}{\sqrt{d_k}} + \boldsymbol{B}\right) \cdot \boldsymbol{V}_1^i \\ \boldsymbol{g_2head}^i = \text{softmax}\left(\dfrac{\boldsymbol{K}_2^i \cdot (\boldsymbol{Q}_1^i)^\intercal}{\sqrt{d_k}} + \boldsymbol{B}\right) \cdot \boldsymbol{V}_2^i \end{cases} \tag{13}$$

where $\boldsymbol{K}_1^i$, $\boldsymbol{V}_1^i$, and $\boldsymbol{Q}_1^i$ are different projections of $\boldsymbol{S_1}$. Meanwhile, $\boldsymbol{K}_2^i$, $\boldsymbol{V}_2^i$, and $\boldsymbol{Q}_2^i$ represent different linear projections of $\boldsymbol{S_2}$. If the number of heads h is 1, we will abbreviate $\boldsymbol{K}_1^i$, $\boldsymbol{V}_1^i$, and $\boldsymbol{Q}_1^i$ in Eq. (13) as $\boldsymbol{K}_1$, $\boldsymbol{V}_1$, and $\boldsymbol{Q}_1$, respectively. Correspondingly, for $h = 1$, $\boldsymbol{K}_2^i$, $\boldsymbol{V}_2^i$, and $\boldsymbol{Q}_2^i$ in Eq. (13) are abbreviated as $\boldsymbol{K}_2$, $\boldsymbol{V}_2$, and $\boldsymbol{Q}_2$, respectively.

From Eq. (13) both the $\boldsymbol{I_1}$- and $\boldsymbol{I_2}$- guided attention mechanisms take the association between modalities into consideration. After inserting Eq. (13) into Eq. (11) and (12), we could get the output of W-MG$_1$A and W-MG$_2$A (i.e., $\boldsymbol{S}_{12}^f$ and $\boldsymbol{S}_{21}^f$) in Fig. 3.

Note that despite the subtle differences between SW-MG$_1$A and SW-MG$_2$A in window partitioning and cross-window attention, both adopt guided attention mechanisms to model long-range dependencies. Therefore, the outputs of SW-MG$_1$A and SW-MG$_2$A, i.e., $\boldsymbol{S}_{12}^{f'}$ and $\boldsymbol{S}_{21}^{f'}$ in Fig. 3, are attainable by substituting Eq. (13) into Eq. (11) and (12), respectively.

C. Hierarchical Fusion and Reconstruction

Multimodal image fusion aims to explore and utilize the inter-modal useful information to compensate for the deficiency of single-source images and yield high-quality fused images.For example, to fully leverage low-level spatial information and high-level structure information among multi-scale features, a bi-directional hierarchical feature collaboration (BHFC) module is given in [53]. [54] gives a hierarchical multimodal fusion architecture to explore multiple bidirectional translation processes, thereby generating dual multimodal fusion embeddings. Motivated by this, we propose to build a hierarchical fusion branch involving a shallow interactive fusion module (SIF), a multi-level interactive fusion module (MIF), and a high-level interactive fusion and reconstruction module (HRM) to fully integrate the multi-modality, multi-granularity, and multi-view features. As depicted in Fig. 2, the former concentrates on the fusion of dual-stream low-level features. Meanwhile, the latter resorts to aggregating dual-stream local and long-range context semantics from various levels.

1) Shallow Interactive Fusion: Clearly, aligning and integrating dual-stream information is of great significance for effective multi-modal fusion. In this regard, a shallow interactive fusion module (SIF) and a multi-level interactive fusion module (MIF) are built and used in cascade.

SIF aims to fuse the shallow multi-view features from two streams for subsequent processing. Inside the SIF, we first perform concatenation and 3 × 3 convolution operations on the fine resolution features from the GIR$_1$ and GIR$_2$ (i.e., $\boldsymbol{S_f^1}, \boldsymbol{S_f^2}$). After GIR$_1$ and GIR$_2$, $\boldsymbol{S_f^1}$ and $\boldsymbol{S_f^2}$ is obtained with the size of $\frac{H}{4} \times \frac{W}{4} \times (2 * d)$, respectively. Note that multi-head self-attention embodies automatic focus statistics, which could reveal and exploit the significance of each channel. Therefore, we further employ the multi-head self-attention mechanism to adaptively integrate the valuable shallow features from the detail and content streams. The specific calculation process involved in SIF is defined in Eq. (14).

$$\boldsymbol{S_f} = MHead(Conv(Concat(\boldsymbol{S_f^1}, \boldsymbol{S_f^2}))) \tag{14}$$

where $Concat(\cdot, \cdot)$ and $Conv$ stand for the concatenation and 3×3 convolution, respectively. As shown in Fig. 2, $\boldsymbol{S_f^1}$ denotes the output from the front part of GIR$_1$ in the detail stream, $\boldsymbol{S_f^2}$ corresponds to the output from the front part of GIR$_2$ module in the content stream. Then, through SIF, the fused low-level features $\boldsymbol{S_f}$ is captured with the size of $\frac{H}{4} \times \frac{W}{4} \times (4 * d)$.

2) Multi-Level Interactive Fusion: For the sake of further aggregating the deep semantics from two streams, but not forgetting shallow features, a multi-level interactive fusion module (MIF) is constructed. MIF is committed to integrating multi-view features from SIF and deep semantics from the latter part of GIR$_1$ and GIR$_2$. Specifically, since patch merging has the advantage over pooling in terms of information preservation, MIF first utilizes patch merging to obtain lower-scale features from the SIF module (i.e., $\boldsymbol{S_f}$). Then, to absorb multi-level local and long-range contextual information from two modalities (i.e., $\boldsymbol{M_f^1}, \boldsymbol{M_f^2}$), multi-modal features of diverse scales are concatenated followed by 1×1 convolution for feature aggregation. Finally, similar to SIF, the multi-head self-attention mechanism is applied for attentive feature fusion. The expression for MIF is given by

$$\boldsymbol{M_f} = MHead(Conv(Concat(PM(\boldsymbol{S_f}), \boldsymbol{M_f^1}, \boldsymbol{M_f^2}))) \tag{15}$$

where $\boldsymbol{S_f}$ is the output of SIF. PM represents the operation of patch merging. Due to the operation of patch merging, the spatial size of $\boldsymbol{S_f}$ is reduced to $\frac{H}{8} \times \frac{W}{8}$, which is the same as that of $\boldsymbol{M_f^1}$ and $\boldsymbol{M_f^2}$. Then, based on Eq. (15), we could capture the multi-level semantic features $\boldsymbol{M_f} \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times (8*d)}$.

From Eq. (15), it is obvious that MIF not only fuses multi-scale and multi-modal features but also achieves multi-view fusion via multi-head attention.

3) High-Level Interactive Fusion and Reconstruction: To generate high-quality fused images, the core problem is to comprehensively explore and aggregate compatible and credible complementary features. To this end, we further construct a high-level interactive fusion and reconstruction (HRM) branch. This part focuses mainly on hierarchical feature fusion through Swin Transformer blocks. The main reason is that Swin Transformer has excellent long-range modeling capability. Particularly, multi-head self-attention can obtain different perceptions from cross-modal hierarchical features with multi-scales and levels. This is beneficial for comprehensive and multi-view analysis of explored features.

As manifested in Fig. 2, the concatenation and 3 × 3 convolution are utilized for feature aggregation from two streams and the fusion branch. Specifically, the aggregation block takes multi-level $\boldsymbol{M_f}$ and high-level $\boldsymbol{M_f^1}$, $\boldsymbol{M_f^2}$ semantic information from dual-stream as inputs. Then, the aggregated features are injected into 3× Swin Transformer blocks demonstrated as follows.

$$\boldsymbol{D_f} = ST_{3\times}(Conv(Concat(\boldsymbol{M_f}, \boldsymbol{M_f^1}, \boldsymbol{M_f^2}))) \tag{16}$$

where $ST_{3\times}$ represents three consecutive Swin Transformer blocks. Here, Conv means 3×3 convolution. Through Eq. (16), we could acquire the high-level aggregated semantic features $\boldsymbol{D_f} \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times (4*d)}$. There are two reasons for using convolution in aggregation blocks. The first is to employ their local perception ability to perceive detailed information further. The second is to project the multi-modal and multi-level features into one shared space for semantic alignment.

Regarding feature decoding, the patch expansion operation is first leveraged, which first restores the size of the feature map from $\frac{H}{8} \times \frac{W}{8}$ to $\frac{H}{4} \times \frac{W}{4}$. Then, enlightened by residual connection, the second convolution block combines low-level information $\boldsymbol{S_f^1}, \boldsymbol{S_f^2}$ from two streams with the expanded $\boldsymbol{D_f}$. After that, the aggregated features are decoded based on 4 consecutive Swin Transformer blocks, which guarantees that multi-modal features of different scales, levels, and views can be considered. The specific construction process of $\boldsymbol{R_f} \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times (2*d)}$ can be described as

$$\boldsymbol{R_f} = ST_{4\times}(Conv(Concat(PE(\boldsymbol{D_f}), \boldsymbol{S_f^1}, \boldsymbol{S_f^2}))) \tag{17}$$

where PE denotes the patch expansion operation, which could increase the size of the feature maps.

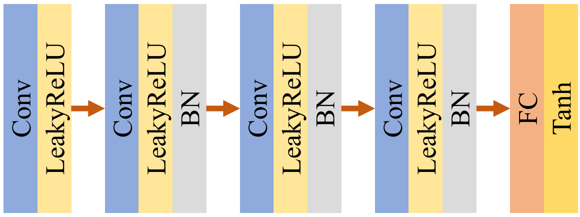


Fig. 4. Architecture of Discriminator

Then, through size expansion and channel reduction, we can reconstruct the $H \times W \times 1$ fused image $\boldsymbol{I_f}$ as

$$\boldsymbol{I_f} = Conv(PE(\boldsymbol{R_f})) \tag{18}$$

where Conv is the 1×1 convolution for channel reduction.

D. Discriminators

For generator optimization in IG-GAN, it is sensible to introduce discriminators for adversarial learning. Specifically, we give two discriminators for the mutual game with the generator. They are designed to classify the generated image and misclassify the original multimodal images. For example. the first one can identify that the first mode image is true, while the fused image is false. The second discriminator is used to discriminate the fused image from the second mode image. Concretely, It can recognize the first modal image as true but the fused image as false.

Concerning structural symmetry, the two discriminators have the same structure but do not share parameters. As described in Fig. 4, each discriminator contains four convolutional modules, which involve Conv (convolution), LeakyReLU, and BN (Batch Normalization) layers. Finally, the fully connected (FC) layer is used, followed by the Tanh activation function.

E. Loss Function

The loss function is of crucial importance to guide the network optimization and boost the mutual game between the generator and discriminators. To promote the training of IG-GAN without supervision, a comprehensive loss function respecting the generator and discriminators is given by

$$\mathcal{L}{=}\mathcal{L}_G + \mathcal{L}_{Dis1} + \mathcal{L}_{Dis2} \tag{19}$$

where $\mathcal{L}_G$ reflects the generator loss, and $\mathcal{L}_{Dis1}$ and $\mathcal{L}_{Dis2}$ correspond to the losses of discriminator$_1$ and discriminator$_2$, respectively.

(1) **Generator Loss**

The core of multimodal fusion is to use the inter-modality complementary information to enrich texture details and maintain content integrity. In this regard, the generator loss is given by

$$\mathcal{L}_G = \lambda_1 \mathcal{L}_{con} + \lambda_2 \mathcal{L}_{ei} + \mathcal{L}_{com} \tag{20}$$

where λ_1 and λ_2 are balancing parameters, the $\mathcal{L}_{con}$, L_{ei}, $\mathcal{L}_{com}$ represent the content loss, edge intensity loss, and compatibility loss, respectively.

• Content Loss: The content loss $\mathcal{L}_{con}$ aims to retain the completeness of spatial structures. Note that the structural similarity index (SSIM) is one of the most widely used indices for image fusion, which reflects the structural similarity between images from three perspectives (i.e., light intensity, contrast, and structure). Therefore, we adopt SSIM to manifest the content completeness of the fused image. Additionally, as the mean square error (MSE) is an intuitive index for measuring the discrepancy between modalities pixel-by-pixel, it is employed for constructing the content loss. The specific content loss involving structural similarity loss $\mathcal{L}_{ssim}$ and mean square error loss $\mathcal{L}_{mse}$ is

$$\mathcal{L}_{con} = a\mathcal{L}_{ssim} + b\mathcal{L}_{mse} \tag{21}$$

where a and b are harmonic coefficients, which are set to 5 and 1 in the experiment, respectively.

$\mathcal{L}_{ssim}$ is defined in Eq. (6). Specifically, considering the need to treat each modality equally, in Eq. (6), both w_1 and $w2$ are set to 1. Eq. (22) gives the definition of $\mathcal{L}_{mse}$.

$$\mathcal{L}_{mse} = w_3\|\boldsymbol{I_f} - \boldsymbol{I_1}\|_2 + w_4\|\boldsymbol{I_f} - \boldsymbol{I_2}\|_2 \tag{22}$$

where $\|\cdot\|_2$ represents the l_2-norm, w_3 and w_4 are trade-off parameters, which are set to 1 in this paper.

• Edge Intensity Loss: It is well known that the richer the texture, the larger the total gradient of the image. Therefore, to enhance the texture details of the fused image, the edge intensity loss is expressed as

$$\mathcal{L}_{ei} = 1 - ((\nabla_x \boldsymbol{I_f})^2 + (\nabla_y \boldsymbol{I_f})^2)^{\frac{1}{2}} \tag{23}$$

where ∇ denotes the gradient operator, x and y represent the directions of the derivative. The above three losses form our generator loss function, which aims to maximize the probability that the fused image comes from both modalities.

In contrast, the discriminator adopts a multi-class classifier to determine that the fused image is neither an infrared image nor a visible image. Then, with continuous adversarial learning, the generator can estimate the probability distribution of both the infrared and visible images, which could generate a fused image with significant contrast and rich texture details.

• Compatibility Loss: The fused image is the complementary fusion result of the multimodal input images. This means that the fused image should be compatible with the input images. Assume the input image is true. Then, the generator should try to maximize the probability that the fused image is still true. Eq. (24) gives the specific formulation of compatibility loss.

$$\mathcal{L}_{com} = \mathbb{E}[\log(1 - D_1(\boldsymbol{I_f}))] + \mathbb{E}[\log(1 - D_2(\boldsymbol{I_f}))] \tag{24}$$

where $\mathbb{E}$ means expectation, $\boldsymbol{I_f}$ represents the fused image, and D_1 and D_2 stand for the discriminator$_1$ and discriminator$_2$, respectively.

(2) **Discriminator Loss**

The primary task of discriminators is to improve the performance of the generator through game confrontation. Note that the compatibility loss helps the generated fused image to be compatible or in fidelity with the input multimodal images. Therefore, the discriminator should recognize the incompatibilities and judge the fused image as false. Eq. (25) gives the discriminator loss $\mathcal{L}_{Dis}$ as follows.

$$\mathcal{L}_{Dis} = \mathbb{E}[-\log D(\boldsymbol{I_s})] + \mathbb{E}[-\log(1 - D(\boldsymbol{I_f}))] \tag{25}$$

where D stands for discriminator, $\boldsymbol{I_s}$ represents for the source image, and $\boldsymbol{I_f}$ denotes the fused image generated by the generator. The loss function described above is used for both source images.

IV. Experiments

This section first describes the experimental settings, e.g., four commonly used multimodal datasets, comparative methods, evaluation metrics, and parameters setting. Secondly, quantitative and qualitative comparisons with several state-of-the-art methods are provided. Finally, the effectiveness of the specific design is assessed by ablation studies.

A. Experimental Settings

1) **Dataset**

The OS dataset consists of fine-resolution optical and SAR images [55]. The optical images were collected via the Google Earth platform, whereas SAR images were captured by the Chinese C-band sensor Gaofen-3, in spotlight mode. The dataset contains 10692 image pairs with size 256 × 256. The TNO dataset is an infrared and visible image data set provided by TNO in the Netherlands. These image pairs include various scenes with size 360 × 270, 505 × 510, and 768 × 576 [56]. The RGB-NIR Scene dataset includes 477 images captured in RGB and near-infrared (NIR) [57]. There are nine scenes in total: country, field, forest, indoor, mountain, old building, street, urban, and water.

2) **Comparative Methods**

To evaluate the effectiveness of our proposed method, fourteen state-of-the-art fusion methods were employed for performance comparison, including DDcGAN [37], FusionGAN [36], GAN-FM [58], GANMcC [12], PMGI [29], RFNNest [28], SDDGAN [32], STDFusionNet [59], U2Fusion [27], SwinFusion [46], TGFuse [45], DetailGAN [50], ATFuse [60], and PSFusion [61]. Among the fourteen methods, DDcGAN, FusionGAN, GAN-FM, GANMCC, SDDGAN, TGfuse, and DetailGAN all achieve multimodal image fusion through the architecture of GAN.

3) **Evaluation Metrics**

Concerning quantitative comparison, seven typical evaluation criteria were adopted. Edge Intensity (EI), Spatial Frequencies (SF), and Average Gradient (AG) mainly attempt to reflect the edge intensity, spatial frequencies, and average gradient of a fused image, respectively. Meanwhile, The sum of the correlations of differences (SCD) and the correlation coefficient (CC) are self-explanatory [62]. Both Visual Information Fidelity (VIF) and Visual information fidelity for fusion (VIFF) were introduced to characterize the visual information fidelity of the images [63], [64].

4) **Settings**

All the experiments were performed on an NVIDIA GeForce GTX 3090 GPU with batch size 24. In the first

TABLE I
QUANTITATIVE COMPARISON WITH STATE-OF-THE-ARTS ON THE OS DATASET.

	EI	**SF**	**SCD**	**VIF**	**AG**	**CC**	**VIFF**
DDcGan [37]	92.974	20.417	1.095	1.261	9.041	0.75	0.511
FusionGAN [36]	48.579	10.3	0.813	0.459	4.611	0.747	0.233
GANFM [58]	126.069	32.792	1.518	1.014	12.896	0.78	0.533
GANMcC [12]	65.644	16.01	1.348	0.602	6.659	0.798	0.405
PMGI [29]	105.072	26.639	1.379	0.752	10.621	0.783	0.485
RFNNest [28]	57.028	11.094	1.393	0.706	5.325	0.784	0.444
SDDGAN [32]	95.341	21.002	1.452	1.144	9.147	0.763	0.682
STDFusionNet [59]	67.365	16.944	1.073	0.633	6.548	0.729	0.198
SwinFusion [46]	91.094	22.927	1.479	0.739	9.186	0.767	0.341
TGFuse [45]	130.072	37.135	1.507	0.953	13.687	0.762	0.542
U2Fusion [27]	123.799	31.13	1.496	0.88	12.515	0.783	0.512
DetailGAN [50]	78.123	16.766	1.386	0.536	7.557	**0.8**	0.422
ATFuse [60]	152.741	47.01	0.833	0.924	16.752	0.729	0.372
PSFusion [61]	122.293	32.817	1.574	0.93	12.918	0.782	0.557
IG-GAN	**165.012**	**48.251**	**1.656**	**2.001**	**17.564**	0.784	**0.765**

25 epochs of the training process, the learning rate was set as 0.002 without discriminators. After that, the learning rate was set to 0.0001. Before feeding the image pairs into our IG-GAN, all images are resized to 256 × 256 in advance. Moreover, all comparison methods are set with reference to their authors.

B. Results on the OS Dataset

1) Quantitative Comparison: To evaluate the effectiveness of IG-GAN, this subsection mainly focuses on a quantitative comparison of the OS dataset. Table I provides the quantitative comparison with 11 popular methods in terms of EI, SF, SCD, VIF, AG, CC, and VIFF on the OS dataset. The optimal and suboptimal results are marked in bold and underlined font respectively.

From Table I we can observe that GANMcC and DetailGAN exhibit superior results respecting CC. This demonstrates the effectiveness of their loss function, which requires the fused image to be consistent with the original inputs. While except CC, our IG-GAN has obvious advantages over the others in terms of EI, SF, SCD, VIF, AG, and VIFF. For example, our EI is 165.012, which is almost 12.3 higher than the next-best comparator and 116 higher than the lowest one. Regarding SF, we achieved 1.24 advantages over the second-placed and 37.95 advantages over the worst-placed. In terms of AG, our method is greater than the third-place TGfuse by nearly 3.9. Concerning VIFF, we achieved 0.765, whereas the worst is 0.198, which is over 0.56 lower. As EI, SF, SCD, VIF, AG, and VIFF characterize the detail texture, integrity, and fidelity of fused images from different views, the obvious advantage in these criteria shows the superiority of our fusion method regarding detail richness and content integrity.

2) Qualitative comparison: For illustrative purposes, Fig. 5 shows an example of the fused images generated by each method. The first two images in this figure correspond to the original optical (OPT) and SAR images, respectively. For better display, an enlarged view of the object area from the red line frame is presented.

From Fig. 5 we can see that balancing the completeness and clarity of the objects is a challenge for most fusion methods. For example, respecting DDcGAN, FusionGAN, PMGI, RFNNest, DetailGAN, and PSFusion, the integrity of the objects is acceptable. In comparison, the fused images are blurred and lack precise edge contours. By contrast, GANMcC, SwinFusion, U2Fusion, and ATFuse yield clearer outlines of the objects. Unfortunately, the completeness and contrast are not satisfactory. In view of GAN-FM, STD-fusionNet, TGFuse, and PSFusion, their fused images are superior. While in terms of the brightness and saliency of the objects, they are still some way behind the results of IG-GAN. Apparently, our IG-GAN can give complete objects, clear contours, and obvious contrast.

C. Results on the TNO Dataset

1) Quantitative Comparison: This subsection provides the comparative experiments corresponding to different methods conducted on the TNO dataset. Table II shows the specific experimental results respecting seven criteria. Bold and underlined results represent the optimal and suboptimal results, respectively.

As illustrated in Table II, ATFuse performs excellently on the TNO dataset. It can acquire suboptimal fusion results for EI, SF, and AG. This means that the fused images are rich in structural and detailed information. Note that DDcGAN and GANMcC achieve the suboptimal VIF and CC, respectively. This demonstrates that they are good at preserving the original input information. Regarding IG-GAN, instead of performing well in some single aspects, it is superior to others on the whole. For example, in terms of EI, SF, SCD, VIF, AG, and VIFF, it always performs the best. Even for CC, it is also the third-best only after DetailGAN and GANMcC. This reveals that IG-GAN is advantageous to producing a high-

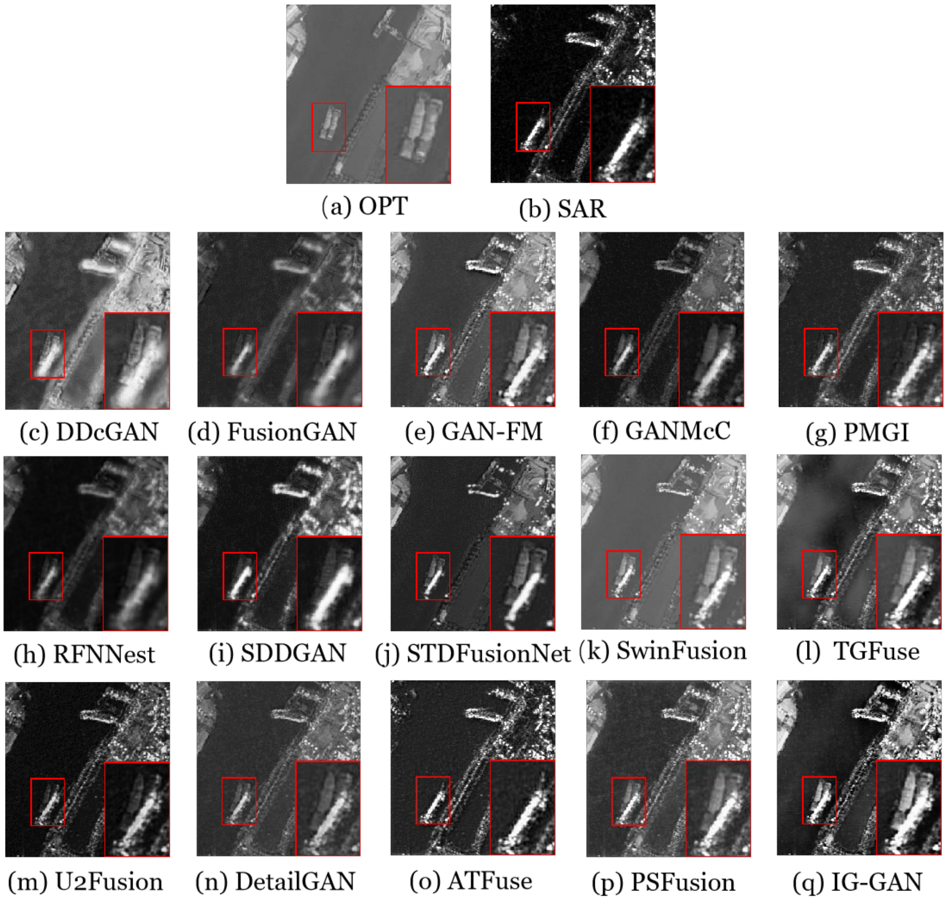


Fig. 5. Qualitative Comparison on the OS dataset.

quality fused image involving great texture, complete content, and high-fidelity vision.

2) Qualitative Comparison: To be more intuitive, Fig. 6 further depicts an example of the original images and the corresponding fused images generated by the fourteen methods. The first two images in this figure correspond to the infrared (IR) and visible (VIS) images, respectively. For better display, an enlarged view of the object area from the red line frame is provided

As depicted in Fig. 6, constrained by illumination, only weak objects are challenging to recognize in the visible images. In contrast, the infrared image can present the objects with deficient details. In this scenario, image fusion should prioritize the infrared image over the visible light image. Otherwise, as observed in DDcGAN and FusionGAN, the visible light image significantly influences the fused image. Consequently, the objects in the fused image become highly blurred, as indicated in Table II. As demonstrated in Fig. 7, IG-GAN tends to outperform others for all test images. In alignment with Fig. 7, our IG-GAN can produce high-quality fusion images with complete content, a clear outline, and strong contrast.

D. Results on the RGB-NIR Scene Dataset

1) Quantitative Comparison: To evaluate the effectiveness of IG-GAN, this subsection further provides a quantitative comparison conducted on the RGB-NIR Scene dataset. Table III shows the corresponding results respecting EI, SF, SCD, DF, AG, and VIFF. Bold and underlined represent the optimal and suboptimal results, respectively.

From Table III we can observe that IG-GAN still has excellent performance. Despite not being the best for all criteria, it can achieve the top two for these criteria. For example, in terms of EI, SCD, DF, and VIFF, it is suboptimal and quite close to the optimal. Regarding SF, our method is superior to the others. For AG, though IG-GAN is the third-best place, it is closely different from the optimal 8.911 and suboptimal 8.856 by 0.08 and 0.025, respectively. The outstanding performance on four popular datasets demonstrates that IG-GAN achieves the preservation of inter-modal consistency and the comprehensive fusion of complementary information.

2) Qualitative Comparison: To be more intuitive, Fig. 8 further gives an example of the original images and the corresponding fused images generated by fourteen methods. In this figure, the first two images correspond to the near-infrared (NIR) and visible (VIS) images, respectively. For better display, we also provide an enlarged view of the red object area.

As shown in Fig. 8, some methods integrate the texture and brightness from VIS and NIR well. While respecting edge sharpness and visual contrast, our IG-GAN still outperforms the others. Specifically, the tree is clearly described with

TABLE II
QUANTITATIVE COMPARISON WITH SEVERAL STATE-OF-THE-ART METHODS ON THE TNO DATASET.

	EI	**SF**	**SCD**	**VIF**	**AG**	**CC**	**VIFF**
DDcGAN [37]	44.486	11.193	1.479	**1.191**	4.550	0.702	0.612
FusionGAN [36]	24.142	6.240	1.037	0.524	2.417	**0.727**	0.258
GAN-FM [58]	46.763	12.526	1.545	1.038	4.846	0.682	0.494
GANMcC [12]	25.895	6.139	1.347	0.615	2.546	0.683	0.422
PMGI [29]	36.832	8.749	1.528	0.885	3.606	0.704	0.543
RFNNest [28]	28.644	5.873	1.568	0.729	2.682	0.701	0.513
SDDGAN [32]	36.799	8.991	1.556	1.159	3.608	0.676	0.658
STDFusionNet [59]	44.111	11.807	1.361	1.188	4.456	0.659	0.473
SwinFusion [46]	41.237	10.639	1.71	0.757	4.142	0.681	0.472
TGFuse [45]	41.748	11.044	1.498	0.857	4.232	0.670	0.482
U2Fusion [27]	51.445	11.861	1.607	1.083	5.060	0.700	0.683
DetailGAN [50]	22.994	5.15	1.567	0.386	2.134	0.718	0.3
ATFuse [60]	50.587	12.512	1.5516	0.747	4.885	0.668	0.36
PSFusion [61]	53.311	12.926	1.613	0.85	**7.173**	0.687	0.494
IG-GAN	**55.225**	**13.903**	**1.697**	1.177	5.418	0.704	**0.717**

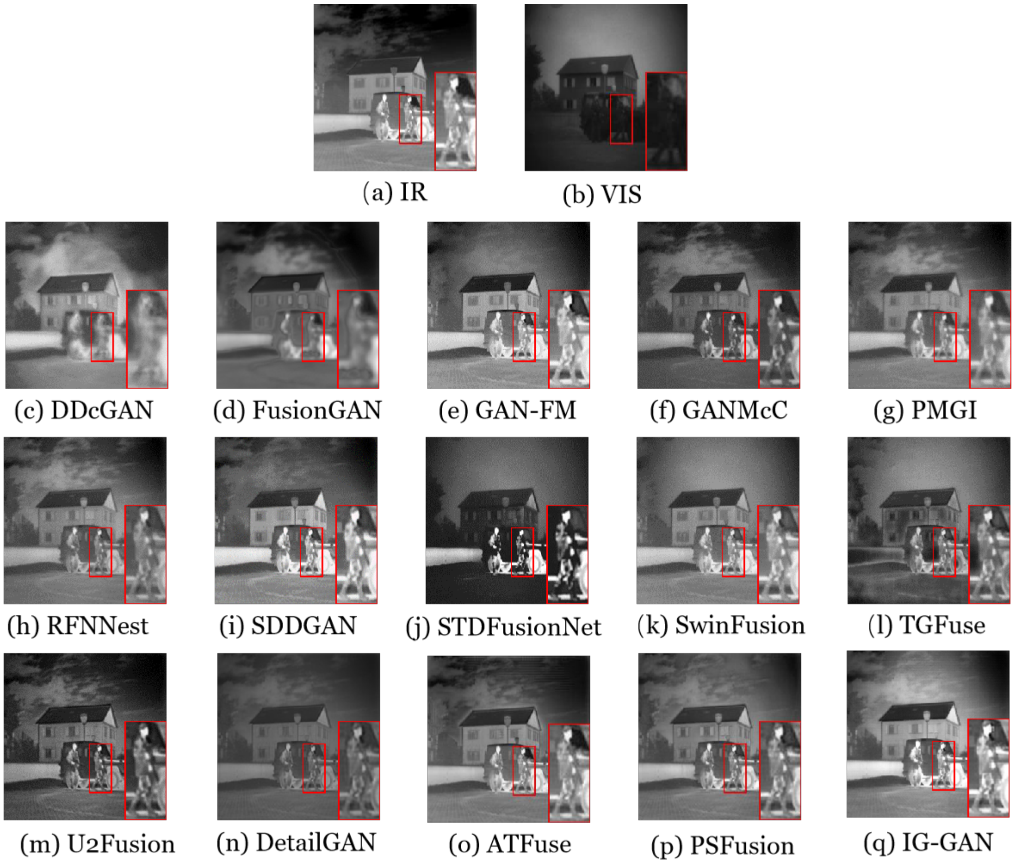
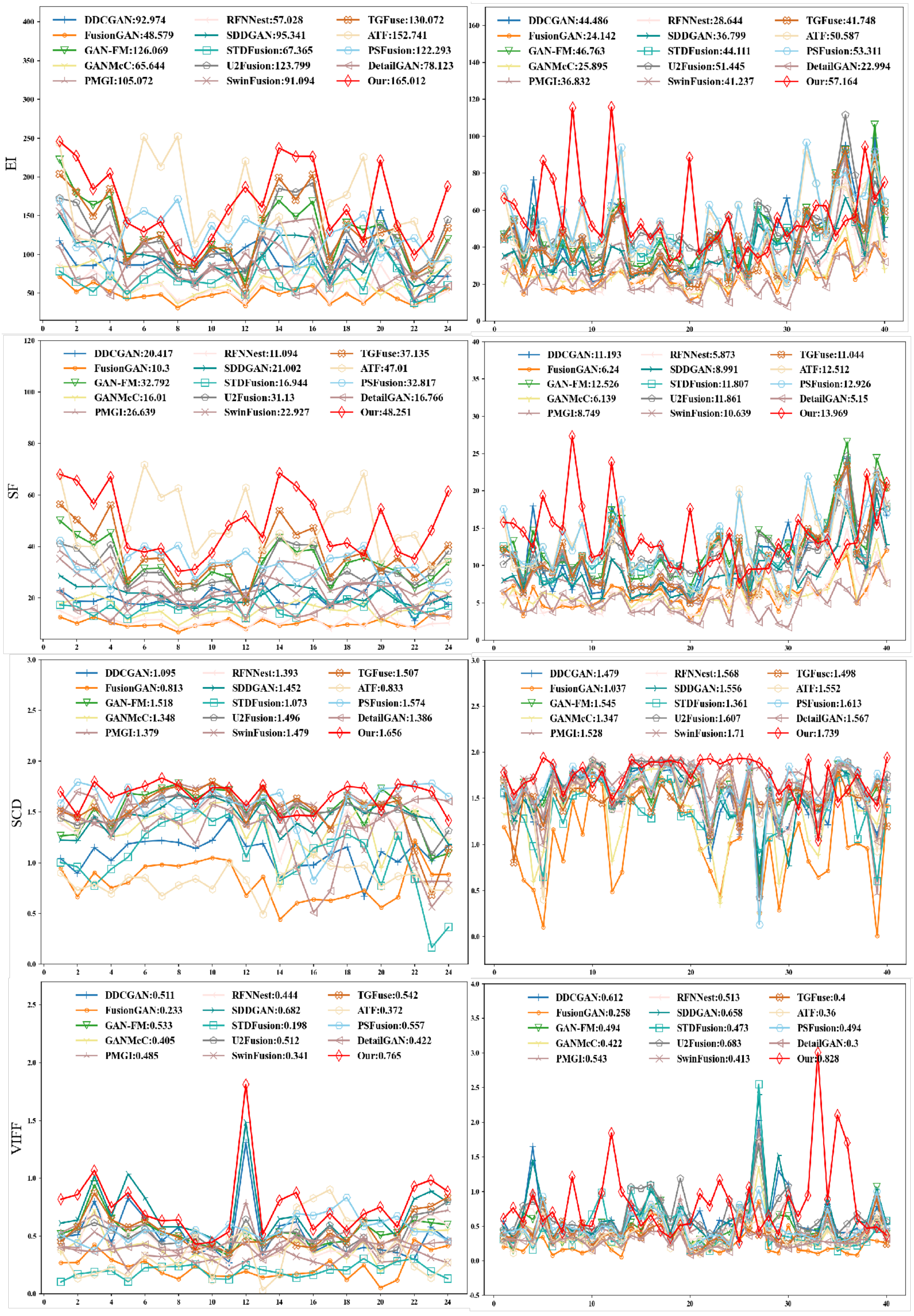

(a) IR
(b) VIS
(c) DDcGAN
(d) FusionGAN
(e) GAN-FM
(f) GANMcC
(g) PMGI
(h) RFNNest
(i) SDDGAN
(j) STDFusionNet
(k) SwinFusion
(l) TGFuse
(m) U2Fusion
(n) DetailGAN
(o) ATFuse
(p) PSFusion
(q) IG-GAN

Fig. 6. Quantitative Comparison on the TNO Dataset.

branches and leaves that stand out and have the highest contrast and structural information. This reflects its unique advantages in details and content exploration and cross-modal concordance enhancement.

E. Computational Complexity and Efficiency

Note that model parameter quantity (Params), floating-point operations per second (FLOPs), and inference time are three important metrics for measuring the complexity and efficiency of deep learning algorithms. Therefore, to evaluate the complexity and efficiency of our IG-GAN, we further give the comparison between IG-GAN and four typical multi-modal fusion methods regarding params, FLOPs, and Inference time.When the input image size is 256×256,



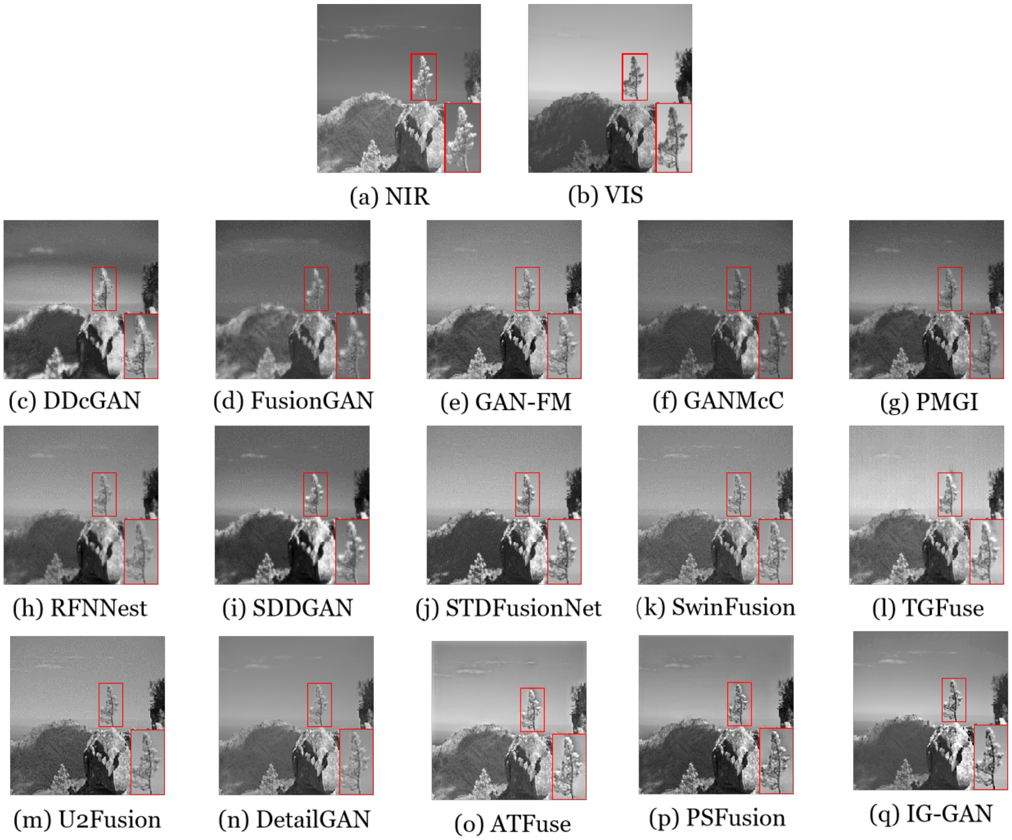
(a) OS dataset

(b) TNO dataset

Fig. 7. Performance comparison curves with fourteen deep fusion methods on the three datasets: (a) OS dataset, (b) TNO dataset.

Table IV presents a comparative analysis of model parameters and computational complexity between SwinFusion, TGFuse, ATFuse, PSFusion, and IG-GAN in terms of Params, FLOPs, and inference time.

There are three reasons for taking SwinFusion, TGFuse, ATFuse, and PSFusion for complexity and efficiency comparison. The first reason is based on the architectural similarities between SwinFusion and IG-GAN, both of which



(a) NIR (b) VIS

(c) DDcGAN (d) FusionGAN (e) GAN-FM (f) GANMcC (g) PMGI

(h) RFNNest (i) SDDGAN (j) STDFusionNet (k) SwinFusion (l) TGFuse

(m) U2Fusion (n) DetailGAN (o) ATFuse (p) PSFusion (q) IG-GAN

Fig. 8. Qualitative comparison on the RGB-NIR dataset.

TABLE III
QUANTITATIVE COMPARISON ON THE RGB-NIR SCENE DATASET.

	EI	SF	SCD	DF	AG	VIFF
DDcGAN [37]	70.778	19.593	0.709	9.611	7.236	0.700
FusionGAN [36]	46.960	12.275	0.330	5.978	4.683	0.427
GAN-FM [58]	81.528	23.374	1.020	11.347	8.417	0.708
GANMcC [12]	52.968	13.981	0.697	6.773	5.310	0.519
PMGI [29]	67.805	17.491	0.828	8.643	6.736	0.629
RFNNest [28]	45.661	9.834	1.158	4.832	4.294	0.582
SDDGAN [32]	54.594	13.389	0.925	6.678	5.319	0.711
STDFusionNet [59]	78.242	19.202	**1.358**	9.050	7.575	**0.879**
SwinFusion [46]	72.477	19.752	1.140	9.712	7.433	0.650
TGFuse [45]	72.904	20.300	1.114	9.657	7.421	0.685
U2Fusion [27]	**89.776**	22.339	1.237	11.142	**8.911**	0.772
DetailGAN [50]	55.34	12.524	1.098	6.118	5.302	0.567
ATFuse [60]	83.192	22.385	1.086	11.283	8.575	0.591
PSFusion [61]	85.687	23.241	1.203	**11.676**	8.856	0.728
IG-GAN	86.445	**23.941**	1.318	11.608	8.831	0.863

are hybrid fusion networks combining CNN and Transformer. In this regard, it is necessary to compare the complexity and efficiency between SwinFusion and IG-GAN. The second reason is that, as given in Table I, Table II, and Table III, TGFuse is an effective GAN-based fusion method, which also involves the Transformer and CNN. Hence, besides the effectiveness, we further provide the complexity and efficiency comparison between IG-GAN and TGFuse. The third reason is that ATFuse and PSFusion are highly competitive fusion methods proposed in the past two years. Regarding this, besides the quality of the fused image, we should further compare the complexity and complexity of IG-GAN with ATFuse and PSFusion. Table IV presents a comparative analysis of model parameters and computational complexity between SwinFusion, TGFuse, ATFuse, PSFusion, and IG-GAN in terms of Params, FLOPs, and inference time.

TABLE IV
MODEL PARAMETER COMPARISON.

Model	Input	Params	FLOPs	Inference Time
SwinFusion	(256 × 256)	3.895M	63.731G	1.035s
TGFuse	(256 × 256)	549.359M	15.945G	0.229s
ATFuse	(256 × 256)	262.939K	5.405G	0.213s
PSFusion	(256 × 256)	45.899M	1.234T	1.031s
IG-GAN	(256 × 256)	3.538M	15.904G	0.246s

According to Table IV, although the number of model parameters for IG-GAN is not the least, it exhibits low model complexity and high fusion efficiency. In specific, IG-GAN has the second lowest number of model parameters after ATFuse. The PSFusion has a significantly higher model size and computational complexity compared to IG-GAN. Specifically, PSFusion's model size is almost 13 times larger than IG-GAN, leading to a corresponding increase in FLOPs, reaching 1.234T compared to IG-GAN's 15.904G. This highlights the trade-off between fusion performance and computational complexity in PSFusion. This reveals the trade-off between the fusion performance and computational complexity in PSFusion. Compared to SwinFusion, the model size of IG-GAN is reduced to over 75%, which leads to a $0.357M$ decrease in FLOPs. Additionally, IG-GAN's inference time is less than a quarter of SwinFusion's. This demonstrates that IG-GAN has a lower model complexity and higher fusion efficiency compared to SwinFusion. Concerning TGFuse, it is a generative adversarial fusion network consisting of a spatial Transformer and a channel Transformer. Despite that TGFuse has less inference time than IG-GAN, it relies on high model complexity, with model parameters exceeding 155 times that of IG-GAN. Therefore, Table IV indicates that although IG-GAN is not the lightest fusion method, it has a relatively small model complexity and high algorithm efficiency.

F. Ablation Studies

The excellent performance of IG-GAN relies on our well-designed structure and loss function. To this end, an experimental ablation study of our proposed shallow interactive fusion module (SIF) and multi-level interactive fusion module (MIF) is first analyzed. Then, we further give an ablation analysis concerning our loss function w.r.t. $\mathcal{L}_{\rceil\rangle}$ on the OS dataset.

1) Ablation Study of Shallow and Multi-Level Interactive Fusion Modules: The ablation experiments were carried out under four situations as shown in Table V. For clarity, we use the blue font next to the up arrow to indicate the increment.

From Table V we find that the fusion performance is slightly improved after employing SIF alone. When MIF alone is introduced, there is greater improvement in terms of all criteria. While for IG-GAN involving both SIF and MIF, the best performance is attained for most criteria. The main reason for this is that they are conducive to the multi-grained, multi-level, and multi-view integration of dual-stream features.

2) Ablation study of the Edge Intensity loss function: To assess the importance of edge intensity loss $\mathcal{L}_{ei}$, comparative experiments without $\mathcal{L}_{ei}$ were conducted in the same settings as our original IG-GAN. Experimental results are depicted in Table VI. For clarity, we adopt the blue font next to the up arrow to indicate the increment.

As shown in Table VI, despite there being a slight and negligible decline for VIF and CC, the employment of $\mathcal{L}_{ei}$ exhibits an evident performance advantage for the other criteria.

G. Application to Object Detection

To explore the application potential of IG-GAN to multi-modal object detection, experiments on two public RGB-IR object detection datasets: (1) DroneVehicle dataset [65] and (2) FLIR dataset [66], are conducted.

1) Dataset: **(1) DroneVehicle:** The DroneVehicle dataset is a large-scale vehicle detection dataset based on UAV aerial photography, containing both visible and infrared modalities. The dataset covers a full range of lighting environments and has many different occlusion information with variations in image scale and shooting angle. The dataset contains five categories, i.e., ”car”, ”van”, ”bus”, ”truck”, and ”freight car”. In this dataset, 17990 pairs of images are used for training, whereas 8980 pairs of images are used for testing. Additionally, all the images are resized to 640×640.

(2) FLIR: The aligned FLIR dataset [66] is an autopilot dataset taken on city streets and highways and contains both daytime and nighttime lighting conditions. This dataset consists of three categories. In our experiments, we leverage 4113 pairs of images for training, whereas 515 pairs of images are used for detection. Similar to the DroneVehicle dataset, all the images are resized to 640×640.

2) Metrics: Mean average precision (mAP) denotes the mean of average precision (AP) across all categories, which is given by

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{26}$$

where AP_i represents the average precision (AP) of the i^{th} category.

TABLE V
ABLATION STUDY ON SHALLOW AND MULTI-LEVEL INTERACTIVE FUSION MODULES.

SIF	MIF	EI	SF	SCD	VIF	AG	CC	VIFF
✘	✘	147.678	43.010	1.633	1.682	15.670	0.784	0.696
✘	✔	157.242	45.263	1.655	1.894	16.684	**0.788** ($\uparrow \mathbf{0.004}$)	**0.788** ($\uparrow \mathbf{0.092}$)
✔	✘	147.781	43.033	1.610	1.694	15.599	0.776	0.696
✔	✔	**165.012** ($\uparrow \mathbf{17.334}$)	**48.251** ($\uparrow \mathbf{5.241}$)	**1.656** ($\uparrow \mathbf{0.023}$)	**2.001** ($\uparrow \mathbf{0.319}$)	**17.564** ($\uparrow \mathbf{5.241}$)	0.784	0.765 ($\uparrow \mathbf{0.069}$)

TABLE VI
ABLATION STUDY OF THE EDGE INTENSITY LOSS.

$\mathcal{L}_{ei}$	EI	SF	SCD	VIF	AG	CC	VIFF
✔	**165.012** ($\uparrow \mathbf{11.398}$)	**48.251** ($\uparrow \mathbf{3.409}$)	1.656	**2.001** ($\uparrow \mathbf{0.104}$)	**17.564** ($\uparrow \mathbf{1.186}$)	0.784	**0.765** ($\uparrow \mathbf{0.030}$)
✘	153.614	44.842	**1.659**	1.897	16.378	**0.789**	0.735

TABLE VII
COMPARISONS OF PERFORMANCES ON THE DRONEVEHICEL DATASET IN TERMS OF MAP50, MAP75, AND MAP50:95.

Method	DroneVehicle			FLIR		
	mAP50	mAP75	mAP50:95	mAP50	mAP75	mAP50:95
YOLOv5	0.757	0.51	0.467	0.802	0.368	0.409
IG-GAN + YOLOv5	**0.828**($\uparrow \mathbf{0.71}$)	**0.721**($\uparrow \mathbf{0.211}$)	**0.598**($\uparrow \mathbf{0.131}$)	**0.852**($\uparrow \mathbf{0.05}$)	**0.454**($\uparrow \mathbf{0.86}$)	**0.461**($\uparrow \mathbf{0.53}$)

Concerning the positional accuracy of the detection box, mAP50, mAP75, and mAP50:95 are widely used, which reflect the mAP values under different Intersection Over Union (IOU) thresholds. Concretely, mAP50 and mAP75 provide the mAP when the IOU threshold is 0.5 and 0.75, respectively. Respecting mAP50:95, it is the average of the mAP values at IOU thresholds ranging from 0.50 to 0.95 in steps of 0.05.

3) Experimental Setting and Results: To evaluate the effectiveness of IG-GAN for enhancing object detection performance, experiments are conducted on the DroneVehicle dataset and the FLIR dataset.

In specific, IG-GAN is first applied to fuse the visible and infrared image pairs and generate the fused images. Then, the fused images are fed into the detection model for training and testing. Note that YOLOv5 [67] is a popular lightweight object detection model owing to its efficiency and excellent detection performance. Regarding this, YOLOv5 https://github.com/ultralytics/yolov5 is trained and tested based on the single modal images and the multimodal fused images through IG-GAN, respectively.

Table VII lists the detection accuracy of images before and after IG-GAN fusion respecting mAP50, mAP75, and mAP50:95. From Table VII we can find that the IG-GAN + YOLOv5 method, significantly outperforms the YOLOv5 method on both the DroneVehicle and FLIR datasets across three metrics (mAP50, mAP75, and mAP90). For example, on the DroneVehicle dataset, the mAP50, mAP75, and mAP90 are 0.757, 0.51, and 0.467, respectively. After IG-GAN, the corresponding detection accuracies are improved by 0.71, 0.211, and 0.131, respectively. On the FLIR dataset, the integration of IG-GAN with YOLOv5 also boosts performance across multiple metrics. Fig. 9 depicts the four pairs of detection results on the DroneVehicle dataset with and without IG-GAN fusion.

In Fig. 9, the first column shows the ground truth of the four scenes. Meanwhile, the second and the third column gives the detection results of YOLOv5 corresponding to images with and without IG-GAN fusion. The red boxes indicate that the objects are missing detection. From Fig. 9 we can see that IG-GAN is beneficial to reduce missed detection by integrating RGB and infrared information. For example, under low light and extremely dark conditions, it can enhance the detection performance by leveraging the infrared information. This demonstrates the potential of IG-GAN to enhance object detection tasks, particularly in challenging scenarios by leveraging multimodal complementary information.

V. CONCLUSION

This paper describes a guided dual-stream progressive interactive generative adversarial fusion network for multi-modal images (IG-GAN). In this network, the details and content streams are first established with mutual collaboration rather than independently, which contributes to detail and content exploration and cross-modal concordance enhancement. Specifically, guided interactive running-in modules (GIR_1, GIR_2) are developed within a dual stream for inter-modal alignment, cooperation, and enhancement. Then, for multi-level dual-stream information fusion, a shallow interactive fusion module (SIF) followed by a multi-level interactive fusion module (MIF) is built. Concerning fine decoding and fused image generation, a high-level interactive fusion and reconstruction module (HRM) is further constructed. This is beneficial to integrate multi-level local-global contextual information. Additionally, for the

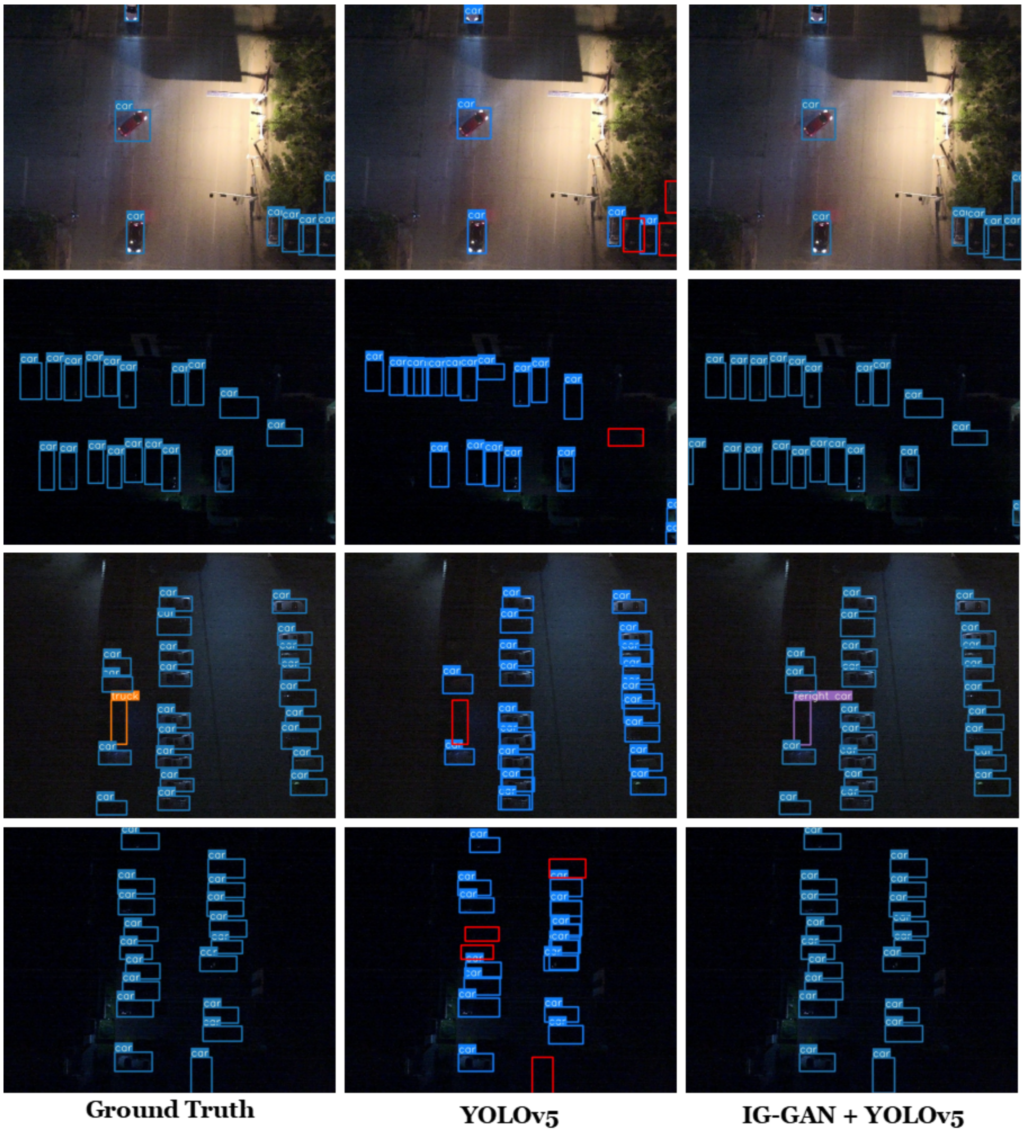


Fig. 9. Visualization of groundtruth and detection results respecting YOLOv5, and combination of IG-GAN and YOLOv5. The red boxes indicate that the objects are missed detection.

sake of network optimization without supervision, we further provide an objected loss function facilitating the generation of complete and detailed fusion images.

Comparative experiments with fourteen state-of-the-art deep fusion methods were conducted on OS, TNO, and RGB-NIR Scene Datasets. Quantitative experimental results show that although many methods are highly effective, IG-GAN exhibits an evident advantage over the other fourteen methods in texture details. Consistent with quantitative comparison, the fused images show that IG-GAN has superiority in completeness, texture details, and contrast. Additionally, an ablation study was performed concerning SIF, MIF, and $\mathcal{L}_{ei}$. The results of the ablation experiments show that these components play a crucial role in improving the fusion performance of IG-GAN.

It is worth noting that a primary objective of Digital Twins (DTs) is to maintain coherence across multiple datasets, such as aligning point data with image data. In this context, IG-GAN holds promise for exploring consistency within the DTs' framework in the future.

However, most existing fusion methods, including IG-GAN, rely on the registered multimodal data pairs for exploring and fusing inter-modal complementary information. Therefore, for the non-paired multimodal data, how to explore and enhance the cross-modal consistency information and then achieving complementary information fusion remains a challenge.

In addition, to embed IG-GAN into multimodal object detection or tracking models, it is crucial to reduce the complexity of the fusion model and improve fusion efficiency. Note that Mamba is a simpler, more efficient, and flexible architecture compared to Transformer. In this regard, we will strive to introduce Mamab to build more lightweight fusion networks in the future, thereby boosting the application potential of IG-GAN for downstream tasks, e.g., object detection and tracking.

References

[1] D. Hong, C. Li, B. Zhang, N. Yokoya, J. A. Benediktsson, and J. Chanussot, “Multimodal artificial intelligence foundation models: Unleashing the power of remote sensing big data in earth observation,” *The Innovation Geoscience*, vol. 2, no. 1, p. 100055, 2024.

[2] D. Hong, J. Hu, J. Yao, J. Chanussot, and X. X. Zhu, “Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 178, pp. 68–80, 2021.

[3] S. Salcedo-Sanz, P. Ghamisi, M. Piles, M. Werner, L. Cuadra, A. Moreno-Martínez, E. Izquierdo-Verdiguier, J. Muñoz-Marí, A. Mosavi, and G. Camps-Valls, “Machine learning information fusion in earth observation: A comprehensive review of methods, applications and data sources,” *Information Fusion*, vol. 63, pp. 256–272, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253520303171

[4] D. Hong, B. Zhang, H. Li, Y. Li, J. Yao, C. Li, M. Werner, J. Chanussot, A. Zipf, and X. X. Zhu, “Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks,” *Remote Sensing of Environment*, vol. 299, p. 113856, 2023.

[5] M. Chen, X. Wang, H. Wang, and S. Zhao, “A uav-based energy-efficient and real-time object detection system with multi-source image fusion,” *Journal of Circuits, Systems and Computers*, vol. 31, no. 09, p. 2250166, 2022.

[6] C. Li, B. Zhang, D. Hong, J. Yao, and J. Chanussot, “Lrr-net: An interpretable deep unfolding network for hyperspectral anomaly detection,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–12, 2023.

[7] X. He, Y. Chen, L. Huang, D. Hong, and Q. Du, “Foundation model-based multimodal remote sensing data classification,” *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[8] Y. Fu and X. Wu, “A dual-branch network for infrared and visible image fusion,” *CoRR*, 2021.

[9] D. Hong, J. Yao, C. Li, D. Meng, N. Yokoya, and J. Chanussot, “Decoupled-and-coupled networks: Self-supervised hyperspectral image super-resolution with subpixel fusion,” *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[10] Poulain, V., Inglada, J., Spigai, M., Tourneret, J.-Y., Marthon, and P., “High-resolution optical and sar image fusion for building database updating,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 8, pp. 2900–2910, 2011.

[11] C. Li, B. Zhang, D. Hong, J. Zhou, G. Vivone, S. Li, and J. Chanussot, “Casformer: Cascaded transformers for fusion-aware computational hyperspectral imaging,” *Information Fusion*, p. 102408, 2024.

[12] 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 Transactions on Instrumentation and Measurement*, vol. PP, no. 99, pp. 1–1, 2020.

[13] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, “Beyond bilinear: Generalized multi-modal factorized high-order pooling for visual question answering,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. PP, no. 99, pp. 1–13, 2017.

[14] S. Li, X. Kang, L. Fang, J. Hu, and H. Yin, “Pixel-level image fusion: A survey of the state of the art,” *Information Fusion*, vol. 33, 2017.

[15] B. K. S. Kumar, “Image fusion based on pixel significance using cross bilateral filter,” *Signal,Image & Video Processing*, 2015.

[16] Y. Liu, X. Chen, R. K. Ward, and Z. J. Wang, “Image fusion with convolutional sparse representation,” *IEEE Signal Processing Letters*, no. 99, pp. 1–1, 2016.

[17] D. P. Bavirisetti, “Multi-sensor image fusion based on fourth order partial differential equations,” in *20th International Conference on Information Fusion (Fusion), 2017*, 2017.

[18] H. Li and X. Wu, “Infrared and visible image fusion using latent low-rank representation,” *CoRR*, 2018.

[19] M. Zhou, J. Huang, K. Yan, D. Hong, X. Jia, J. Chanussot, and C. Li, “A general spatial-frequency learning framework for multimodal image fusion,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[20] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, “Image fusion meets deep learning: A survey and perspective,” *Information Fusion*, vol. 76, no. 11, 2021.

[21] J. Ma, C. Chen, C. Li, and J. Huang, “Infrared and visible image fusion via gradient transfer and total variation minimization,” *Information Fusion*, vol. 31, pp. 100–109, 2016.

[22] P. Li, “Didfuse: deep image decomposition for infrared and visible image fusion,” in *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 2021, pp. 976–976.

[23] K. R. Shahi, P. Ghamisi, B. Rasti, P. Scheunders, and R. Gloaguen, “Unsupervised data fusion with deeper perspective: A novel multisensor deep clustering algorithm,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 284–296, 2022.

[24] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, “Pansharpening by convolutional neural networks,” *Remote Sensing*, vol. 8, no. 7, p. 594, 2016.

[25] Y. Zheng, J. Li, Y. Li, J. Guo, X. Wu, and J. Chanussot, “Hyperspectral pansharpening using deep prior and dual attention residual network,” *IEEE transactions on geoscience and remote sensing*, vol. 58, no. 11, pp. 8059–8076, 2020.

[26] J. Yao, B. Zhang, C. Li, D. Hong, and J. Chanussot, “Extended vision transformer (exvit) for land use and land cover classification: A multimodal deep learning framework,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[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*, vol. 44, 07 2020.

[28] L. A. Hui, A. Xjw, and B. Jk, “Rfn-nest: An end-to-end residual fusion network for infrared and visible images,” *Information Fusion*, 2021.

[29] 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,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 12 797–12 804, 04 2020.

[30] H. Xu, J. Ma, and X. P. Zhang, “Mef-gan: Multi-exposure image fusion via generative adversarial networks,” *IEEE Transactions on Image Processing*, vol. PP, no. 99, pp. 1–1, 2020.

[31] W. Liao, Q. Zhang, B. Yuan, G. Zhang, and J. Lu, “Heterogeneous multidomain recommender system through adversarial learning,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 11, pp. 8965–8977, 2022.

[32] H. Zhou, W. Wu, Y. Zhang, J. Ma, and H. Ling, “Semantic-supervised infrared and visible image fusion via a dual-discriminator generative adversarial network,” *IEEE Transactions on Multimedia*, pp. 1–1, 11 2021.

[33] H. Zhang, Y. Song, C. Han, and L. Zhang, “Remote sensing image spatiotemporal fusion using a generative adversarial network,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 5, pp. 4273–4286, 2021.

[34] J. Gao, Q. Yuan, J. Li, H. Zhang, and X. Su, “Cloud removal with fusion of high resolution optical and sar images using generative adversarial networks,” *Remote Sensing*, vol. 12, no. 1, p. 191, 2020.

[35] M. Zhou, J. Huang, D. Hong, F. Zhao, C. Li, and J. Chanussot, “Rethinking pan-sharpening in closed-loop regularization,” *IEEE transactions on neural networks and learning systems*, 2023.

[36] J. Ma, Y. Wei, P. Liang, L. Chang, and J. Jiang, “Fusiongan: A generative adversarial network for infrared and visible image fusion,” *Information Fusion*, vol. 48, pp. 11–26, 2019.

[37] 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 Transactions on Image Processing*, vol. 29, pp. 4980–4995, 2020.

[38] D. Hong, B. Zhang, X. Li, Y. Li, C. Li, J. Yao, N. Yokoya, H. Li, P. Ghamisi, X. Jia, A. Plaza, P. Gamba, J. A. Benediktsson, and J. Chanussot, “Spectralgpt: Spectral remote sensing foundation model,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5227–5244, 2024, dOI:10.1109/TPAMI.2024.3362475.

[39] V. VS, J. M. J. Valanarasu, P. Oza, and V. M. Patel, “Image fusion transformer,” *CoRR*, vol. abs/2107.09011, 2021.

[40] J.-F. Hu, T.-Z. Huang, L.-J. Deng, H.-X. Dou, D. Hong, and G. Vivone, “Fusformer: A transformer-based fusion network for hyperspectral image super-resolution,” *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[41] W. G. C. Bandara and V. M. Patel, “Hypertransformer: A textural and spectral feature fusion transformer for pansharpening,” *arXiv preprint arXiv:2203.02503*, 2022.

[42] W. Liu, Y. Lin, W. Liu, Y. Yu, and J. Li, “An attention-based multiscale transformer network for remote sensing image change detection,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 202, pp. 599–609, 2023.

[43] J. Feng, Q. Wang, G. Zhang, X. Jia, and J. Yin, “Cat: Center attention transformer with stratified spatial-spectral token for hyperspectral image

classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[44] C. Li, B. Zhang, D. Hong, X. Jia, A. Plaza, and J. Chanussot, "Learning disentangled priors for hyperspectral anomaly detection: A coupling model-driven and data-driven paradigm," *IEEE Transactions on Neural Networks and Learning Systems*, 2024, dOI: 10.1109/TNNLS.2024.3401589.

[45] D. Rao, X. Wu, and T. Xu, "Tgfuse: An infrared and visible image fusion approach based on transformer and generative adversarial network," *CoRR*, vol. abs/2201.10147, 2022.

[46] 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, p. 18, 2022.

[47] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *CoRR*, vol. abs/2103.14030, 2021.

[48] A. Prakash, K. Chitta, and A. Geiger, "Multi-modal fusion transformer for end-to-end autonomous driving," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7073–7083, 2021.

[49] X. Yan, S. Z. Gilani, H. Qin, and A. Mian, "Structural similarity loss for learning to fuse multi-focus images," *Sensors*, vol. 20, no. 22, p. 6647, 2020.

[50] 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*, p. 85–98, Feb 2020.

[51] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10 002, 2021.

[52] L. Qu, S. Liu, M. Wang, and Z. Song, "Transmef: A transformer-based multi-exposure image fusion framework using self-supervised multi-task learning," *ArXiv*, 2021.

[53] Z. Zhong, X. Liu, J. Jiang, D. Zhao, Z. Chen, and X. Ji, "High-resolution depth maps imaging via attention-based hierarchical multi-modal fusion," *IEEE Transactions on Image Processing*, vol. 31, pp. 648–663, 2022.

[54] J. Tang, K. Li, X. Jin, A. Cichocki, Q. Zhao, and W. Kong, "CTFN: hierarchical learning for multimodal sentiment analysis using coupled-translation fusion network," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP*, 2021, pp. 5301–5311.

[55] Y. Xiang, R. Tao, F. Wang, and H. You, "Automatic registration of optical and sar images via improved phase congruency," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019.

[56] A. Toet, "The tno multiband image data collection," *Data in brief*, vol. 15, pp. 249–251, 2017.

[57] M. Brown and S. Süsstrunk, "Multi-spectral sift for scene category recognition," in *IEEE Conference on Computer Vision & Pattern Recognition*, 2011.

[58] H. Zhang, J. Yuan, X. Tian, and J. Ma, "Gan-fm: Infrared and visible image fusion using gan with full-scale skip connection and dual markovian discriminators," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 1134–1147, 2021.

[59] J. Ma, T. Linfeng, M. Xu, H. Zhang, and G. Xiao, "Stdfusionnet: An infrared and visible image fusion network based on salient target detection," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–13, 01 2021.

[60] L. Jian, S. Xiong, H. Yan, X. Niu, S. Wu, and D. Zhang, "Rethinking cross-attention for infrared and visible image fusion," *arXiv preprint arXiv:2401.11675*, 2024.

[61] L. Tang, H. Zhang, H. Xu, and J. Ma, "Rethinking the necessity of image fusion in high-level vision tasks: A practical infrared and visible image fusion network based on progressive semantic injection and scene fidelity," *Information Fusion*, vol. 99, p. 101870, 2023.

[62] V. Aslantas and E. Bendes, "A new image quality metric for image fusion: The sum of the correlations of differences," *AEU - International Journal of Electronics and Communications*, vol. 69, no. 12, pp. 1890–1896, 2015.

[63] H. R. Sheikh and A. C. Bovik, "A visual information fidelity approach to video quality assessment," vol. 7, no. 2, pp. 2117–2128, 2005.

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

[65] Y. Sun, B. Cao, P. Zhu, and Q. Hu, "Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning," *IEEE Transactions on Circuits and Systems for Video Technology*, p. 6700–6713, Oct 2022.

[66] H. Zhang, E. Fromont, S. Lefevre, and B. Avignon, "Multispectral fusion for object detection with cyclic fuse-and-refine blocks," in *2020 IEEE International conference on image processing (ICIP)*, 2020, pp. 276–280.

[67] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proceedings of the IEEE*, vol. 111, no. 3, pp. 257–276, 2023.