SwinFusion: Cross-domain Long-range Learning for General Image Fusion via Swin Transformer

Jiayi Ma, *Senior Member, IEEE*, Linfeng Tang, Fan Fan, Jun Huang, Xiaoguang Mei, and Yong Ma

***Abstract*—This study proposes a novel general image fusion framework based on cross-domain long-range learning and Swin Transformer, termed as SwinFusion. On the one hand, an attention-guided cross-domain module is devised to achieve sufficient integration of complementary information and global interaction. More specifically, the proposed method involves an intra-domain fusion unit based on self-attention and an inter-domain fusion unit based on cross-attention, which mine and integrate long dependencies within the same domain and across domains. Through long-range dependency modeling, the network is able to fully implement domain-specific information extraction and cross-domain complementary information integration as well as maintaining the appropriate apparent intensity from a global perspective. In particular, we introduce the shifted windows mechanism into the self-attention and cross-attention, which allows our model to receive images with arbitrary sizes. On the other hand, the multi-scene image fusion problems are generalized to a unified framework with structure maintenance, detail preservation, and proper intensity control. Moreover, an elaborate loss function, consisting of SSIM loss, texture loss, and intensity loss, drives the network to preserve abundant texture details and structural information, as well as presenting optimal apparent intensity. Extensive experiments on both multi-modal image fusion and digital photography image fusion demonstrate the superiority of our SwinFusion compared to the state-of-the-art unified image fusion algorithms and task-specific alternatives. Implementation code and pre-trained weights can be accessed at https://github.com/Linfeng-Tang/SwinFusion.**

***Index Terms*—Cross-domain long-range learning, image fusion, Swin transformer.**

I. Introduction

OWING to the limitations of hardware devices, information captured by a single-type sensor or with a single shooting setting cannot comprehensively characterize the imaging scenario [1]. On the one hand, different types of sensors usually capture specific information from multiple perspectives. For instance, the infrared sensor gathers thermal radiation information, which emphasizes prominent targets. The visible sensor generates digital images with abundant texture details by catching the reflected light information [2]. The near-infrared sensor could capture the complementary details that may be lost in the visible images [3]. Moreover, in the field of medical imaging, structural systems (*e.g.*, magnetic resonance imaging (MRI) and computed tomography (CT)) generally offer structural and anatomical information [4]. By contrast, functional systems such as positron emission tomography (PET) could provide functional information on blood flow and metabolic changes [5]. On the other hand, the sensors with different shooting settings usually acquire limited information from the imaging scenario. More specifically, cameras with varied ISO and exposure times only capture information within the dynamic range and inevitably miss information outside the dynamic range. Similarly, the cameras with specific focal lengths only capture the objects within the depth-of-field (DOF) [6]. It is worth mentioning that images captured by different sensors or under multiple shooting settings generally contain complementary information, which encourages us to incorporate these complementary characteristics into a single image. Therefore, the image fusion technique was born. In terms of the difference in imaging devices, image fusion could be divided into multi-modal image fusion and digital photographic image fusion. A schematic illustration of these two types of image fusion scenarios is exhibited in Fig. 1. A single fused image with better scene representation and visual perception facilitates subsequent practical visual applications, such as object detection, tracking, semantic segmentation, scene understanding, *etc.* [7]–[9].

In the past decades, numerous image fusion techniques have been proposed, which can be broadly divided into two categories, *i.e.*, task-specific image fusion schemes [11]–[14] and general image fusion algorithms [10], [15], [16]. Both task-specific image fusion and general image fusion can be further specified into four classes, including traditional framework [17]–[19], convolution neural network (CNN)-based framework [20], [21], auto-encoder (AE)-based framework [22], [23], and generative adversarial network (GAN)-based framework [24]–[26]. Although the frameworks mentioned above can generate considerable fused results, none of them can sufficiently mine and integrate global context both within and across domains. In particular, we assume that images shot by different sensors or under multiple optical settings belong to different domains in this paper. On the one hand, the traditional frameworks usually implement complementary information aggregation in the spatial domain [17] or transform domain [19], [27], but neither of them can exchange information between non-adjacent pixels. Therefore, the traditional framework fails to

Manuscript received April 4, 2022; revised May 25, 2022; accepted June 10, 2022. This work was supported by the National Natural Science Foundation of China (62075169, 62003247, 62061160370), and the Key Research and Development Program of Hubei Province (2020BAB113). Recommended by Associate Editor Qing-Long Han. *(Corresponding author: Yong Ma.)*

Citation: J. Ma, L. F. Tang, F. Fan, J. Huang, X. G. Mei, and Y. Ma, "SwinFusion: Cross-domain long-range learning for general image fusion via Swin transformer," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 7, pp. 1200−1217, Jul. 2022.

The authors are with the Electronic Information School, Wuhan University, Wuhan 430072, China (e-mail: jyma2010@gmail.com; linfeng0419@gmail.com; fanfan@whu.edu.cn; junhwong@whu.edu.cn; meixiaoguang@gmail.com; mayong@whu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JAS.2022.105686

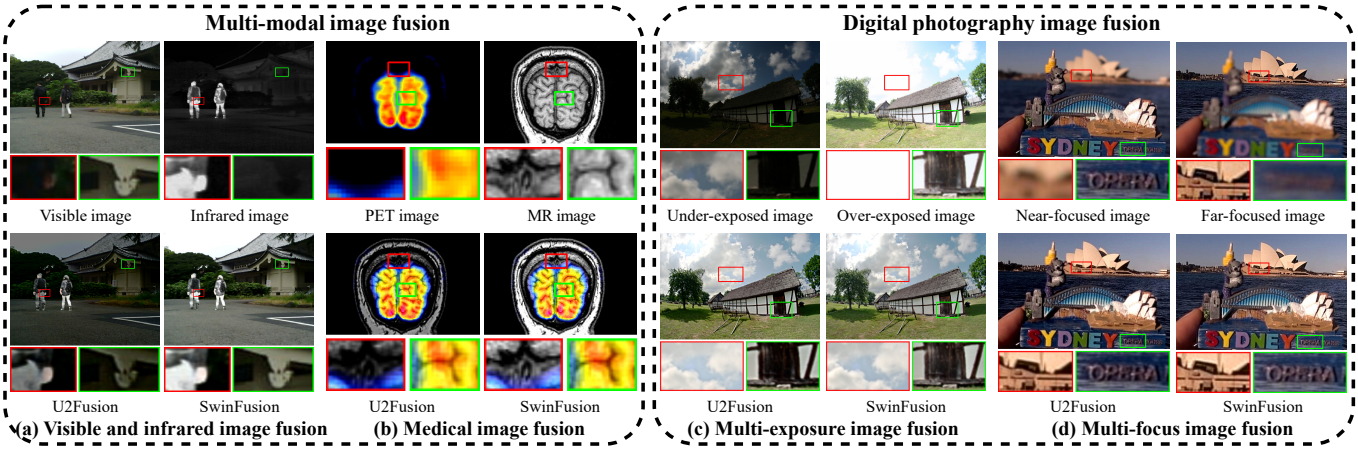


Fig. 1. Schematic illustration of multi-modal image fusion and digital photography image fusion. First row: source image pairs, second row: fused results of state-of-the-art general fusion algorithm, *i.e.*, U2Fusion [10] and our SwinFusion.

perceive the global environment. On the other hand, the basic components of CNN-, AE-, and GAN-based frameworks are convolutional layers that can only mine interactions within the receptive field. However, while exploiting local information for image fusion, these frameworks cannot leverage intra- or inter-domain long-range dependencies to further improve the fused results.

As an alternative to CNN, Transformer [28] devises a self-attention mechanism to capture global interactions between contexts and shows promising performance in several vision problems [29]–[33]. In particular, the image fusion community also introduces Transformer to model the inter-domain long-range dependence and provides competitive fused results [34]–[37]. Nevertheless, there are still some drawbacks that need to be addressed. First, the existing Transformer-based methods merely explore intra-domain interactions, but fail to integrate cross-domain contexts, which is essential for image fusion tasks. Second, vision Transformers for image fusion usually request the input image that can be reshaped to a fixed size (*e.g.*, 256 $\times$ 256), which leads to distorted scenarios in the fused images. Third, the existing fusion Transformers are devised for specific fusion scenarios without considering the intrinsic connection between different fusion tasks.

To address the challenges mentioned above, we devise a general image fusion framework based on cross-domain long-range learning and Swin Transformer for multi-modal image fusion and digital photography image fusion. Our design is primarily developed from the following aspects. On the one hand, we model all image fusion scenarios as structure maintenance, texture preservation, and appropriate intensity control. In particular, we unify the form of loss function that consists of SSIM loss, texture loss, and intensity loss, for all fusion problems. All sub-loss terms follow the same modeling manners of different fusion tasks except for intensity loss, which is tailored to the specific fusion mission for more appropriate intensity perceptions. On the other hand, we design a joint CNN-Transformer image fusion framework to mine the local and global dependencies in source images adequately. The CNN-based shallow feature extraction unit mines the local information in source images. The Transformer-based deep feature extraction unit explores the global interactions among shallow features and generates deep features containing high-level semantic information. Then, the elaborate attention-guided cross-domain fusion module effectively integrates intra- and inter-domain interactions in the deep features. Specifically, the intra-domain fusion unit aggregates the global context in the same domain via the self-attention mechanism. The inter-domain fusion unit models the long-range dependencies between multiple source images and achieves global feature fusion via exchanging the query, key, and value from different domains. Finally, the Transformer-based deep feature reconstruction unit and the CNN-based fused image reconstruction unit leverage global and local information to reconstruct the fused image with superior visual perceptions. It is worth remarking that both self-attention and cross-attention are implemented by the shifted window mechanism (*i.e.*, Swin Transformer [38]), which allows our framework to handle input images with arbitrary size. To sum up, the major contributions of this work can be summarized as follows:

- We propose a joint CNN-Transformer fusion framework for multi-modal image fusion and digital photography image fusion. The proposed framework can sufficiently exploit both local and global information to achieve better complementary characteristics integration.
- A self-attention-based intra-domain fusion unit and a cross-attention-based inter-domain fusion unit are devised to model and integrate the long-range dependencies within the same domain and across domains, respectively.
- Both multi-modal image fusion and digital photography image fusion are generalized to structure maintenance, texture preservation, and appropriate intensity control. Especially, a unified loss function form is defined to constrain all image fusion problems.
- Extensive experiments demonstrate the superiority of our framework compared to state-of-the-art task-specific and general fusion algorithms on both multi-modal image fusion and digital photography image fusion.

The organizational structure of this paper is as follows.

Section II summarizes some relevant research to the proposed framework, including task-specific image fusion, general image fusion, and vision Transformer. Section III provides a detailed discussion of our SwinFusion. In Section IV, we present some qualitative and quantitative results on multi-modal image fusion and digital photography image fusion, as well as performing the ablation study to verify the effectiveness of specific designs. Some concluding remarks are given in Section V.

II. Related Work

The image fusion and vision Transformer are two of the most relevant techniques to our method, here we review some representative research to introduce their developments.

A. Task-specific Image Fusion Methods

As an essential image enhancement technique, image fusion has continued to attract increasing attention in recent years. The mainstream image fusion schemes, especially for task-specific image fusion, can be classified into the following four types of frameworks.

Traditional Image Fusion Framework: Traditional fusion frameworks usually realize image fusion in the spatial domain and transform domain. On the one hand, integrating pixel-level information in the spatial domain is one of the major genres of traditional image fusion. GTF [17] defines infrared and visible image fusion as overall intensity maintenance and texture structure preservation in the spatial domain, and yields the fused image by optimizing the objective function. Awad *et al.* developed an adaptive near-infrared and visible fusion scheme in the spatial domain for visible image detail enhancement [3]. Moreover, Liu *et al.* designed a convolutional sparsity model based on morphological component analysis (CS-MCA) to achieve medical image fusion at the pixel level [39]. They also introduced the local feature descriptor (*i.e.*, Dense SIFT) into the multi-focus image fusion task to perform activity level measurement and match the misregistered pixels between different source images [40]. On the other hand, researchers have also tried to map source images to the transform domain with relevant mathematical transformation and manually design fusion rules in the transform domain to achieve image fusion. Ma *et al.* employed a structural path decomposition technique to transform source images into three conceptually independent components, *i.e.*, signal strength, signal structure, and mean intensity [41]. Then, the multi-exposure image fusion is implemented by merging these three components separately. Furthermore, Li *et al.* proposed a transform domain-based multi-focus image fusion algorithm by combining the sparse feature matrix decomposition and morphological filtering technique [42].

CNN-based Image Fusion Framework: In recent years, convolution neural network (CNN) has gradually become the primary workhorse for image fusion, and has exhibited significant advantages. One form of CNN participating in image fusion employs the pre-trained networks to realize activity level measurement and generate a weight map for hand-crafted features [5], [43]. But the whole fusion process is still based on the traditional fusion framework, such as the Laplace pyramid [5] and guided filtering [43]. Another type of CNN-based image fusion framework is utilizing CNN to learn a direct mapping between source images and fused images (or focus map) in an end-to-end manner [2], [44]. Various researches integrated task-specific prior information into the CNN-based framework to design loss functions and network structures. Specifically, Ma *et al.* proposed an α-matte boundary defocus model to precisely simulate the defocus spread effect and generate realistic data for the training of the multi-focus image fusion network [45]. In order to tackle the difficulty of the blur level estimation around the focused/defocused boundary, Li *et al.* introduced the deep regression pair learning to directly convert the whole image into the binary mask without any patch operation [46]. Zhao *et al.* presented a depth-distilled multi-focus image fusion method by taking depth cues into consideration [47]. They also focused on diversity of features to improve fusion performance [48], [49]. In addition, Han *et al.* devised a deep perceptual enhancement network for multi-exposure image fusion, which contains two separate modules for gathering content details and correcting color distortion, respectively [50]. For visible and infrared image fusion, Long *et al.* designed an aggregated residual dense network that combines the structural advantages of ResNet and DenseNet on the basis of CNN [51]. Moreover, SeAFusion [7] incorporates semantic constraints into the modeling of image fusion for the first time and proposes a gradient residual dense block to boost the description ability for fine-grained details.

AE-based Image Fusion Framework: Contemporaneously, researchers have also explored the auto-encoder-based image fusion framework. Specifically, an auto-encoder pre-trained on large-scale datasets is employed as the feature extractor and image reconstructor, and then a specialized fusion strategy is designed for deep features to implement image fusion. DeepFuse [13] is the pioneer of such fusion frameworks. Afterwards, Li *et al.* introduced the dense connection [22] and nest connection [52], [53] to reinforce the feature extraction capability of the encoder. Moreover, Jian *et al.* injected the attention mechanism into the AE-based fusion framework to reinforce the salient features extracted by the encoder [54]. In order to extract features with greater interpretability, Xu *et al.* tailored the disentanglement representation to the AE-based fusion framework [11]. Nevertheless, all aforementioned methods adopt hand-crafted fusion strategies, *e.g.*, element-wise addition [13], element-wise weight summation [22], and element-wise maximum [20], to merge deep features, which obstruct the fusion models from achieving their optimal performance. For this purpose, Xu *et al.* devised a learnable fusion rule based on pixel-wise classification saliency and interpretable importance evaluation [23].

GAN-based Image Fusion Framework: Generative adversarial network (GAN) can effectively model data distributions even without supervised information, which coincides with the image fusion tasks. Ma *et al.* instructively defined the image fusion problem as a game between the generator and discriminator. Then, they applied GAN to a range of fusion tasks, such as infrared and visible image fusion [55], multi-exposure image fusion [25], multi-focus image fusion [56], and pan-sharpening [57]. However, a single discriminator

fails to take into account the data distributions of multiple domains. Hence, Xu *et al.* proposed the dual-discriminator conditional generative adversarial network (DDcGAN), which leverages two discriminators to constrain the distribution of fused results. Subsequently, Hung *et al.* devised a multi-generator multi-discriminator conditional generative adversarial network (MGMDcGAN) for medical image fusion [26]. Moreover, Li *et al.* injected the multi-scale attention mechanism into the GAN-based fusion framework to encourage the generator and discriminator to pay more attention to the meaningful regions [58], [59].

B. General Image Fusion Methods

Task-specific fusion algorithms are able to exploit relevant priors to improve fusion performance, whereas they ignore the intrinsic associations between different image fusion tasks. Thus, a growing number of researchers are dedicated to developing the unified image fusion framework. MST-SR is the first general image fusion framework that implements complementary information aggregation by combining multi-scale transform (MST) and sparse representation (SR) techniques [15]. Subsequently, Zhang *et al.* [20] designed the first convolution neural network for general image fusion with reference to DeepFuse [13]. In addition, PMGI [16] regards different image fusion problems as proportional maintenance of gradient and intensity, as well as designing a unified form of loss functions. On the basis of PMGI, Zhang *et al.* proposed a squeeze-and-decomposition network and an adaptive decision block to further improve the fusion performance [60]. Moreover, Zhao *et al.* developed a general framework for multi-realm image fusion via learning domain-specific and domain-general feature representations [61]. In particular, considering that different fusion scenarios can promote one another, Xu *et al.* developed a unified unsupervised image fusion model for multi-fusion tasks by combining the learnable information measurement and elastic weight consolidation [10], [62].

It is worth emphasizing that neither task-specific nor general fusion approaches can fully exploit the long-range interactions of images. In other words, these algorithms only merge complementary information from a local perspective but cannot implement global information aggregation.

C. Vision Transformer

Recently, the natural language processing model, *i.e.*, Transformer [28] has received a lot of attention in the computer vision community. There are many Transformer-based models that have achieved impressive performance in various vision tasks, such as visual recognition [29], [63], [64], object detection [30], [65], [66], tracking [67]–[69], segmentation [31], [70], and image restoration [32], [33], [71]. Due to its powerful long-range modeling capability, Transformer has also been introduced to image fusion [34], [35], [37], [72]. Building on the CNN-based fusion framework, VS *et al.* designed a multi-scale fusion strategy based on Spatio-Transformer (*i.e., IFT*), which attends to both local and global contexts [35]. In addition, on the basis of the AE-based fusion framework, Fu *et al.* replaced the CNN architecture with the Patch Pyramid Transformer to extract non-local information from the entire image [37]. However, the auto-encoder only consisting of Transformer fails to effectively extract local information. For this purpose, Zhao *et al.* proposed a sequential DenseNet and dual-transformer architecture, termed DNDT, to extract local and global information, in which dual-transformer reinforces the global information in features before the fusion layer [72]. In addition, Qu *et al.* developed TransMEF [34], which injects parallel Transformer and CNN architecture into the AE-based fusion framework and leverages self-supervised multi-task learning to implement multi-exposure image fusion. Subsequently, Li *et al.* proposed a convolution-guided transformer framework for visible and infrared image fusion (*i.e.*, CGTF), aiming to combine the local feature of CNN and the long-range dependency features of Transformer to generate more satisfactory fused results [73]. Furthermore, Rao *et al.* also introduced Transformer into the GAN-based fusion framework to achieve visible and infrared image fusion [36].

However, the above fusion Transformers merely mine long-range dependencies (or global interactions) from the same domain. In fact, the cross-domain long-range dependencies are more relevant to the image fusion problem. Besides, most of these Transformer-based fusion algorithms, such as IFT [35], DNDT [72], TransMEF [34], and CGTF [73], can only handle input images with fixed size (*e.g.*, 256×256). Moreover, the existing vision Transformers for image fusion only solve specific image fusion problems, but fail to address both multi-modal image fusion and digital photography image fusion scenarios in a unified fusion framework. Thereby, we sufficiently explore the commonalities between different image fusion scenarios. Then, the multi-modal image fusion and digital photography image fusion are modeled uniformly as structure maintenance, texture preservation, and appropriate intensity control. Furthermore, an attention-guided cross-domain fusion module is designed to effectively mine and integrate the intra- and inter-domain global interaction in the fusion process.

III. Methodology

In this section, the multi-modal image fusion and digital photography image fusion are generalized to structure information maintenance, texture detail preservation, and suitable intensity control. We firstly provide the overall framework. Next, the design of the unified loss function is presented.

A. Overall Framework

Let $I_1 \in \mathbb{R}^{H\times W\times C_{in}}$ and $I_2 \in \mathbb{R}^{H\times W\times C_{in}}$ represent two aligned source images from different domains, and $I_f \in \mathbb{R}^{H\times W\times C_{out}}$ is the fused image with complete scene representation. H, W, and C_{in} are the height, width and channel number of input images. C_{out} is the channel number of the fused images. The proposed SwinFusion aims to generate the fused image I_f via merging local and global complementary information in the source images I_1, I_2. As illustrated in Fig. 2, SwinFusion can be divided into three parts: feature extraction, attention-guided cross-domain fusion, and reconstruction.

Feature Extraction: At the beginning, we extract the shallow features F_{SF}^1 and F_{SF}^2 by the multiple convolutional layers

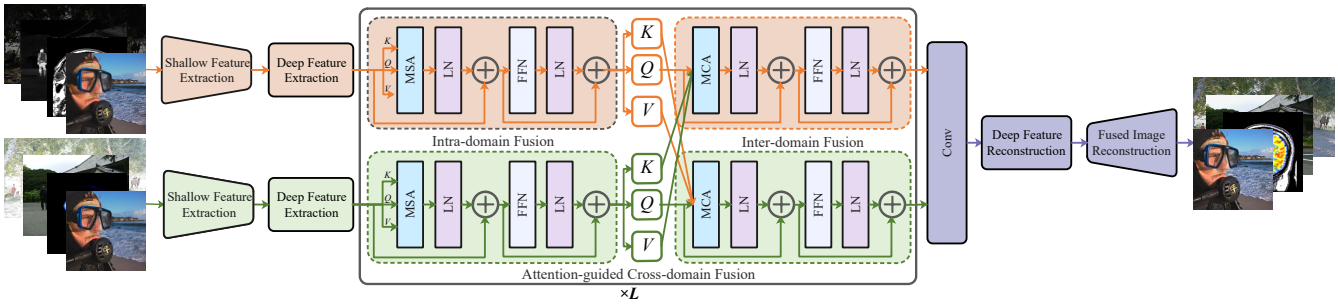


Fig. 2. The framework of the proposed SwinFusion for multi-modal image fusion and digital photography image fusion.

$H_{SE}(\cdot)$ from the source images I_1 and I_2, which can be expressed as:

$$\{F_{SF}^1, F_{SF}^2\} = \{H_{SE}(I_1), H_{SE}(I_2)\}. \tag{1}$$

The convolutional layers are good at early visual processing, resulting in more stable optimization and better results [74]. It also offers a simple yet effective way to extract local semantic information and map this information into the high dimensional feature space. The shallow feature extraction module consists of two convolutional layers with the Leaky Relu activation function, whose kernel size is 3×3, and stride is 1.

After that, we extract deep features F_{DF}^1 and F_{DF}^2 from F_{SF}^1 and F_{SF}^2 as:

$$\{F_{DF}^1, F_{DF}^2\} = \{H_{DE}(F_{SF}^1), H_{DE}(F_{SF}^2)\}, \tag{2}$$

where $H_{DE}(\cdot)$ is the deep feature extraction unit that contains N Swin Transformer layers. The core architecture of the Swin Transformer layer is consistent with the inter-domain fusion unit, which is described in detail below. In this work, N is set to 4.

Attention-guided Cross-domain Fusion: After extracting deep features with sufficient global semantic information, we design an attention-guided cross-domain fusion module (ACFM) to further mine as well as aggregate intra- and inter-domain global context.

First of all, we devise a self-attention-based **intra-domain fusion unit** to effectively integrate the global interactions in the same domain. Shifted window mechanism-based attention is the fundamental component in designing our intra-domain fusion unit. Given features F with the size of $H \times W \times C$, shifted window mechanism first reshapes the input to $\frac{HW}{M^2} \times M^2 \times C$ features via partitioning the input into non-overlapping $M \times M$ local windows, where $\frac{HW}{M^2}$ is the total number of windows. Next, it performs standard self-attention separately for each window. For a local window feature $X \in \mathbb{R}^{M^2 \times C}$, three learnable weight matrices $\mathbf{W}^Q \in \mathbb{R}^{C \times C}$, $\mathbf{W}^K \in \mathbb{R}^{C \times C}$, and $\mathbf{W}^V \in \mathbb{R}^{C \times C}$ that are shared across different windows are employed to project it into the query $\mathbf{Q}$, key $\mathbf{K}$ and value $\mathbf{V}$ by:

$$\{\mathbf{Q}, \mathbf{K}, \mathbf{V}\} = \{X\mathbf{W}^Q, X\mathbf{W}^K, X\mathbf{W}^V\}, \tag{3}$$

Then, the attention function basically computes the dot-product of the query with all keys, which is then normalized with the softmax operator to yield attention scores. The attention mechanism is defined as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} + \mathbf{B})\mathbf{V}, \tag{4}$$

where d_k is the dimension of keys and $\mathbf{B}$ is the learnable relative positional encoding. Referring to the literature [28], we extend the self-attention into multi-head self-attention (MSA) to enable the attention mechanism to consider various attention distributions and make the model capture information from different perspectives. In practice, we perform the attention function for h times in parallel and concatenate the results for multi-head self-attention, in which h is set as 6 in our work. Next, a feed forward network (FFN) that consists of two multi-layer perceptron (MLP) layers with GELU activation layer is deployed to refine the feature tokens yielded by MSA. The layer normalization (LN) is always performed after both MSA and FFN, and the residual connection is applied to both modules. Thus, the full process of intra-domain fusion unit for a local window feature X is formulated as:

$$\begin{aligned} \{\mathbf{Q}, \mathbf{K}, \mathbf{V})\} &= \{X\mathbf{W}^Q, X\mathbf{W}^K, X\mathbf{W}^V\}, \\ \tilde{Z} &= \text{LN}(MSA(\mathbf{Q}, \mathbf{K}, \mathbf{V})) + \mathbf{Q}, \\ Z &= \text{LN}(\text{FFN}(\tilde{Z})) + \tilde{Z}, \end{aligned} \tag{5}$$

where Z is the output of the intra-domain fusion unit with X as the input. The feed forward network (FFN) is as follows:

$$\text{FFN}(X) = \text{GELU}(W_1 X + b_1)W_2 + b_2, \tag{6}$$

where GELU is the gaussian error linear unit. In particular, the Swin Transformer Layer follows the same processing procedure as the intra-domain fusion unit. We also present the framework of two successive Swin Transformer Layers in Fig. 3 to clearly illustrate their processing procedure. It is worth noting that there are no connections across local windows if the partition is fixed for different layers. Thus, following the literature [33], [38], we alternately utilize regular and shifted window partitioning to enable cross-window connections, where shifted window partitioning means shifting the feature by $(\lfloor \frac{M}{2} \rfloor, \lfloor \frac{M}{2} \rfloor)$ pixels before partitioning. Fig. 4 shows an example of the shifted window mechanism for computing attention in the Swin Transformer Layer and intra-domain fusion unit. As can be seen, in layer l, a regular window partitioning scheme is employed, and the attention is computed within each window. In the next layer (*i.e.*, layer $l+1$), the window partitioning is shifted, which results in new windows. Thus, the attention computation

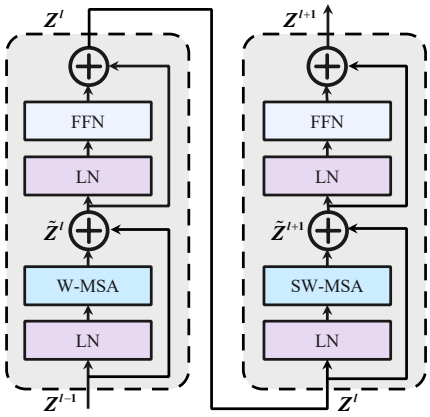


Fig. 3. Two successive Swin Transformer Layers. W-MSA and SW-MSA are multi-head self-attention modules with regular and shifted windowing configurations, respectively.

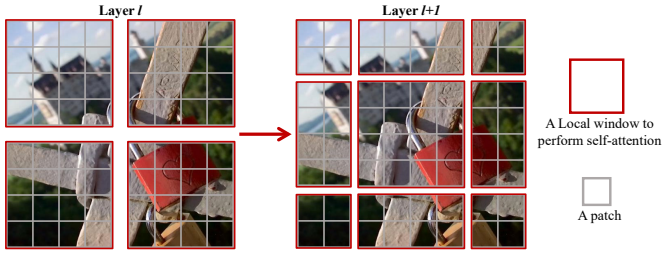


Fig. 4. An illustration of the shifted window mechanism for computing attention in the Swin Transformer Layer, intra-domain fusion unit, and inter-domain fusion unit.

in the new windows crosses the boundaries of the windows in layer l, providing connections among them.

Following the intra-domain fusion unit, we also design a cross-attention-based **inter-domain fusion unit** to further integrate the global interactions between different domains. Both the intra-domain fusion unit and inter-domain fusion unit follow a similar baseline. The principal difference is that the inter-domain fusion unit employs multi-head cross-attention (MCA) instead of MSA to implement global context exchange across domains. Therefore, given two local windows features X_1 and X_2 from different domains, the whole process of the inter-domain fusion unit is defined as:

$$\begin{aligned}
&\{\mathbf{Q}_1, \mathbf{K}_1, \mathbf{V}_2)\} = \{X_1\mathbf{W}_1^Q, X_1\mathbf{W}_1^K, X_1\mathbf{W}_1^V\},\\
&\{\mathbf{Q}_2, \mathbf{K}_2, \mathbf{V}_2)\} = \{X_2\mathbf{W}_2^Q, X_2\mathbf{W}_2^K, X_2\mathbf{W}_2^V\},\\
&\tilde{Z}_1 = \text{LN}(MCA(\mathbf{Q}_1, \mathbf{K}_2, \mathbf{V}_2)) + \mathbf{Q}_1,\\
&\tilde{Z}_2 = \text{LN}(MCA(\mathbf{Q}_2, \mathbf{K}_1, \mathbf{V}_1)) + \mathbf{Q}_2,\\
&Z_1 = \text{LN}(\text{FFN}(\tilde{Z}_1)) + \tilde{Z}_1,\\
&Z_2 = \text{LN}(\text{FFN}(\tilde{Z}_2)) + \tilde{Z}_2.
\end{aligned} \tag{7}$$

As presented in Eq. (7), for $\mathbf{Q}_1$ from domain 1, it incorporates cross-domain information by performing attention weighting with $\mathbf{K}_2$ and $\mathbf{V}_2$ from domain 2, while preserving information in domain 1 through the residual connection and vice versa. Our model deploys L attention-guided cross-domain fusion modules, consisting of cascaded intra-domain fusion unit and inter-domain fusion unit, to alternately integrate the global inter-domain and cross-domain interactions. In order to balance computational efficiency and fusion performance, we set L to 2.

Following the attention-guided cross-domain fusion module, a convolutional layer with spatially invariant filters is deployed to aggregate local information in different domains and enhance the translational equivariance of our SwinFusion, which can be formulated as:

$$F_{FDF} = H_{Conv}(\text{Concat}(F_{AF}^1, F_{AF}^2)), \tag{8}$$

where F_{AF}^1 and F_{AF}^2 represent the output features aggregated by ACFM with F_{DF}^1 and F_{DF}^2 as inputs, respectively. $H_{Conv}(\cdot)$ denotes the convolutional layer with spatially invariant filters and Concat$(\cdot)$ refers to concatenation in the channel dimension. F_{FDF} indicates the fused deep features, which is the input of the feature reconstruction module.

Reconstruction: After fully merging complementary information in different domains, we devise the Transformer-based deep feature reconstruction unit and CNN-based image reconstruction unit to map the fused deep features back to the image space. First of all, the deep feature reconstruction unit $H_{DR}(\cdot)$, containing P Swin Transformer layers, is deployed to refine the fused deep features and restore fused shallow features from a global perspective. This process can be expressed as:

$$F_{FSF} = H_{DR}(F_{FDF}). \tag{9}$$

In order to fully exploit the global context in deep features to recover the fused shallow features, P is set to 4. Then, the CNN-based image reconstruction unit $H_{IR}(\cdot)$ is deployed to reduce the number of channels and generate the fused image I_f, which is denoted as:

$$I_f = H_{IR}(F_{FSF}). \tag{10}$$

The fused image reconstruction unit contains three convolutional layers with kernel size of 3×3 and stride of 1, where the first two layers are followed by the Leaky Relu activation function.

B. Loss Function

In order to model the multi-modal image fusion and digital photography image fusion uniformly, we generalize different image fusion problems into structure maintenance, texture preservation, and suitable intensity control. Accordingly, we design SSIM loss, texture loss, and intensity loss to constrain the network.

SSIM Loss: Considering that the structural similarity (SSIM) index is the most widely used metric, which reflects image distortion from three aspects, *i.e.*, light, contrast, and structure [75], we employ SSIM loss $\mathcal{L}_{ssim}$ to constrain the structural similarity between I_f and I_1, I_2. Specifically, SSIM loss is defined as:

$$\mathcal{L}_{ssim} = w_1 \cdot (1 - ssim(I_f, I_1)) + w_2 \cdot (1 - ssim(I_f, I_2)), \tag{11}$$

where $ssim(\cdot)$ represents the structural similarity operation, which measures the similarity of two images. We consider that both source images have the same contribution to the fused result in terms of structural information. Therefore, we set $w_1 = w_2 = 0.5$ in this work.

Texture Loss: One of the objectives of image fusion is to integrate texture details in the source images into a single fused image. We observe that the texture details in source images could be effectively aggregated by the maximum selection strategy. Thus, texture loss $\mathcal{L}_{text}$, presented in Eq. (12), is designed to guide the network to preserve as many texture details as possible.

$$\mathcal{L}_{text} = \frac{1}{HW} \left\| |\nabla I_f| - \max(|\nabla I_1|, |\nabla I_2|) \right\|_1, \tag{12}$$

where ∇ indicates the Sobel gradient operator, which could measure texture information of an image. $|\cdot|$ stands for the absolute operation, $\|\cdot\|_1$ denotes the l_1-norm, and $\max(\cdot)$ refers to the element-wise maximum selection.

Intensity Loss: An excellent image fusion algorithm is expected to generate a fused image with appropriate intensity according to global apparent intensity information of the source images. For this purpose, we devise the following intensity loss $\mathcal{L}_{int}$ to guide our fusion model to capture proper intensity information:

$$\mathcal{L}_{int} = \frac{1}{HW} \left\| I_f - M(I_1, I_2) \right\|_1, \tag{13}$$

where $M(\cdot)$ is an element-wise aggregation operation, which is associated with the specific fusion scenario. Inspired by IFCNN [20], the element-wise maximum selection, *i.e.*, $\max(\cdot)$ is deployed for visible and infrared image fusion (VIF), medical image fusion (Med), and multi-focus image fusion (MFF). In addition, we leverage the element-wise mean aggregation, *i.e.*, mean$(\cdot)$ for visible and near-infrared image fusion (VIS-NIR) and multi-exposure image fusion (MEF).

Finally, the full objective function for our fusion model is a weighted sum of all sub-loss terms from Eq. (11) to Eq. (13):

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{ssim} + \lambda_2 \mathcal{L}_{text} + \lambda_3 \mathcal{L}_{int}, \tag{14}$$

where λ_1, λ_2, and λ_3 are the hyper-parameters that control the trade-off of each sub-loss term.

IV. Experiments Results and Discussions

In this section, we compare SwinFusion with several state-of-the-art algorithms on both the multi-modal image fusion and digital photography image fusion scenarios by quantitative and qualitative comparisons. We first provide the experimental configurations and then give some implementation details. Subsequently, we conduct quantitative and qualitative comparisons with state-of-the-art alternatives. Extended experiments on other vision tasks are also performed to demonstrate the potential of our method for other computer vision missions. Finally, we verify the effectiveness of specific designs through a series of ablation studies.

A. Experimental Configurations

Datasets: We verify our SwinFusion in multi-modal image fusion and digital photography image fusion. We select three representative scenarios, *i.e.*, visible and infrared image fusion (VIF), visible and near-infrared image fusion (VIS-NIR) as well as medical image fusion (Med) for multi-modal image fusion. Two typical tasks, *i.e.*, multi-exposure image fusion (MEF) and multi-focus image fusion (MFF) are chosen for digital photography image fusion. The training and test data for all fusion tasks are from publicly available datasets. The MSRS dataset [76], [77][1], containing **1,083** training image pairs and **361** testing image pairs, is selected for training and evaluating the visible and infrared image fusion task. We build the training and test datasets based on the publicly available VIS-NIR Scene dataset [78][2] for VIS-NIR. The numbers of the training set and test set are **377** and **100**, respectively. The training and test datasets for medical image fusion are built on the publicly available Harvard medical dataset[3]. Specifically, we select **249** and **20** image pairs for the training and testing of PET and MRI image fusion (Med (PET-MRI)). The numbers of the training set and test set for CT and MRI image fusion (Med (CT-MRI)) are **163** and **20**, respectively. Moreover, the MEF dataset [79][4] is employed to train the model for MEF, and the MEF benchmark dataset [80][5], containing **100** pairs of images with various scenarios, is used as the test set. The MFI-WHU [56][6] and Lytro [81][7] datasets are utilized for training and testing of MFF, respectively, where the Lytro dataset is composed by **20** pairs of color multi-focus images of size 520×520 pixels.

The reasons why we set different test set sizes for different image fusion scenarios are as follows. First of all, some datasets (*e.g.*, MEFB and Lytro) are only adapted to test the performance of different algorithms, so the number of the test sets is the number of the whole datasets, *i.e.*, **100** for the MEFB dataset and **20** for the Lytro dataset. Moreover, the MSRS dataset specifies the number of the test set, *i.e.*, **361**. Because of the number limitation of the Harvard medical dataset, we randomly select **20** test images for all medical image fusion tasks, which is consistent with the Lytro dataset. Besides, **100** test images are randomly selected from the VIS-NIR Scene dataset for the visible and near-infrared image fusion mission, which is kept consistent with the MEFB dataset.

Comparison Algorithms: We select seven state-of-the-art methods, including four general image fusion frameworks and three task-specific approaches as the comparison algorithms for each fusion task. The four unified image fusion algorithms are IFCNN [20], PMGI [16], SDNet [60], and U2Fusion [10]. GTF [17], DenseFuse [22], and FusionGAN [55] are the three task-specific fusion methods for VIF. ANVF [3], DenseFuse [22], and GANMcC [82] are the task-specific comparison algorithms for VIS-NIR. CSMCA [39], EMFusion [4] and DDcGAN [24] are three task-specific approaches selected for medical image fusion tasks. The task-specific alternatives for MEF are SPD-MEF [41], MEFNet [83], and MEF-GAN [25]. SFMD [42], DRPL [46], and MFF-GAN [56] are three comparison methods for MFF. It is worth mentioning that all algorithms, except GTF [17], ANVF [3], CSMCA [39], SPD-MEF [41], and SFMD [42] that are

[1] https://github.com/Linfeng-Tang/MSRS
[2] http://matthewalunbrown.com/nirscene/nirscene.html
[3] http://www.med.harvard.edu/AANLIB/home.html
[4] https://github.com/csjcai/SICE
[5] https://github.com/xingchenzhang/MEFB
[6] https://github.com/HaoZhang1018/MFI-WHU
[7] https://mansournejati.ece.iut.ac.ir/content/lytro-multi-focus-dataset

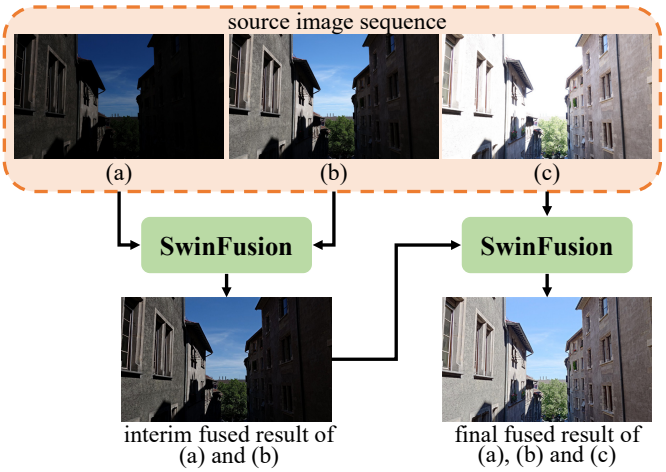


Fig. 5. SwinFusion to fuse multi-exposure image sequence.

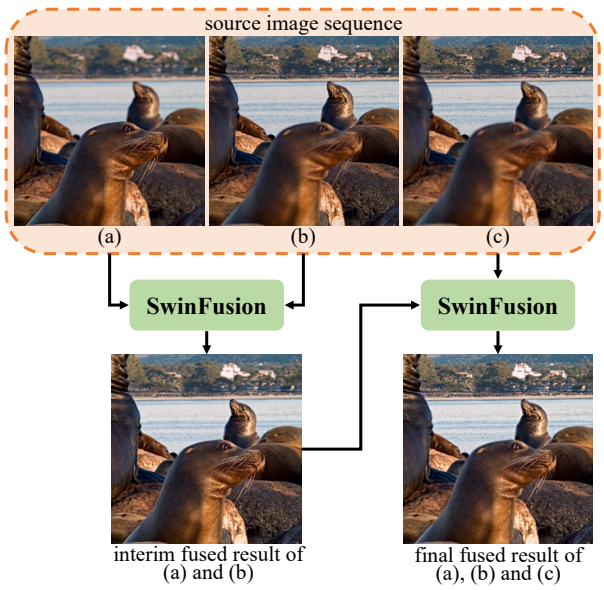


Fig. 6. SwinFusion to fuse multi-focus image sequence.

traditional schemes, are deep learning-based methods.

Evaluation Metrics: Four metrics are selected to quantify the evaluation, including feature mutual information (FMI) [84], Q_{abf}, structural similarity (SSIM) [75], and peak signal-to-noise ratio (PSNR). FMI and Q_{abf} respectively measure the amount of feature information and edge information, which are transferred from source images to the fused image. PSNR reveals the distortion during the fusion process at the pixel level. In addition, SSIM reflects the image distortion from the perspectives of brightness, contrast, and structure. A fusion method with higher FMI, Q_{abf}, SSIM, and PSNR implies better fusion performance.

B. Implementation Details

The batch size is set to 16, and it takes 10 000 training steps for each fusion task. In each step, the images from the training set are randomly cropped into 128×128 patches, which then are normalized to $[0, 1]$. The parameters of our SwinFusion are updated by the Adam optimizer with learning rate initialized to 2×10^{-4} then decayed exponentially. The hyper-parameters that control the trade-off of each sub-loss term are empirically set as $\lambda_1 = 10$, $\lambda_2 = 20$, and $\lambda_3 = 20$. Besides, the window size M is set to 8, referring to SwinIR [33]. The proposed SwinFusion is implemented on the PyTorch platform [85]. Moreover, all experiments are conducted on the NVIDIA TITAN RTX GPU and 2.60GHz Intel(R) Xeon(R) Platinum 8171M CPU.

Dealing with RGB Inputs: RGB inputs are first converted into the YCbCr color space. Next, the Y (luminance) channel is employed as the input of the fusion model since the structural details and intensity information are primarily concentrated in this channel. As for multi-modal image fusion, the fused Y channel is mapped back to the RGB color space along with Cb and Cr (chrominance) channels of the visible image (or PET image) since only the visible image and PET image contain color information. As for digital photography image fusion, the Cb and Cr channels are merged traditionally according to:

$$C_f = \frac{C_1(|C_1 - \tau|) + C_2(|C_2 - \tau|)}{|C_1 - \tau| + |C_2 - \tau|}, \tag{15}$$

where C_1 and C_2 are the Cb or Cr channels of source image I_1 and I_2, respectively. C_f is the fused result of the corresponding channel. τ is set as 128 in this work. Then, the fused Y, Cb, and Cr channels are converted to the RGB color space through the inverse conversion. Thus, both multi-modal image fusion and digital photography image fusion are unified into the single-channel image fusion problem.

Dealing with Sequence Inputs: In practice, a robust framework is expected to fuse sequence images, *i.e.*, more than two images. In this case, we sequentially fuse these source images. The schematic diagrams are presented in Fig. 5 and Fig. 6. As shown in these diagrams, we initially merge two sequence images. Then, the interim result is fused with another source image to generate the final fused image. In this manner, our SwinFusion is theoretically capable of fusing an arbitrary number of sequence images.

C. Results on Multi-modal Image Fusion

Quantitative Comparison: Table I shows the quantitative comparisons between SwinFusion and state-of-the-art algorithms. As one can see, SwinFusion achieves leadership in almost all metrics for multi-modal image fusion tasks. More specifically, the highest FMI and Q_{abf} mean that our approach transfers the most feature and edge information from source images into the fused image. The best SSIM on VIF, VIS-NIR, and Med (PET-MRI) reveals the advantage of structural information maintenance. The proposed framework only lags behind IFCNN by a narrow margin in the SSIM index for Med (CT-MRI). Moreover, our method achieves the best PSNR on VIS-NIR, which implies that our approach has the least information distortion during the fusion process. Although the PSNR of our scheme on VIF lags behind other competitors, this is justifiable. More specifically, our model pays more attention to the salient target regions in the infrared image by adequately integrating the global interactions in source images, resulting in the loss of information in the non-salient areas. A similar

TABLE I
QUANTITATIVE COMPARISON WITH STATE-OF-THE-ARTS IN THE MULTI-MODAL IMAGE FUSION SCENARIOS. RED INDICATES THE BEST RESULT AND BLUE INDICATES THE SECOND BEST RESULT

	VIF								*VIS-NIR*							
	GTF	*DenseFuse*	*FusionGAN*	*IFCNN*	*PMGI*	*SDNet*	*U2Fusion*	*Ours*	*ANVF*	*DenseFuse*	*GANMcC*	*IFCNN*	*PMGI*	*SDNet*	*U2Fusion*	*Ours*
FMI	0.9117	*0.9230*	0.9098	0.9124	0.9021	0.9098	0.9135	**0.9308**	*0.8942*	0.8912	0.8780	0.8797	0.8792	0.8822	0.8805	**0.9012**
Q_{abf}	0.3867	0.4908	0.1393	*0.5246*	0.4019	0.3764	0.4218	**0.6428**	0.6031	*0.6344*	0.5282	0.5785	0.5484	0.5538	0.5499	**0.7046**
SSIM	0.9017	0.9527	0.8182	*0.9572*	0.9254	0.9123	0.9387	**0.9676**	0.9285	*0.9748*	0.9310	0.9616	0.9490	0.9531	0.9521	**0.9859**
PSNR	64.7914	**66.6532**	64.6818	*66.5682*	60.3589	64.8225	66.4867	64.4063	65.4618	66.3679	65.1153	*68.7344*	63.8235	66.4714	66.9182	**68.8832**
	Med (PET-MRI)								*Med (CT-MRI)*							
	CSMCA	*EMFusion*	*DDcGAN*	*IFCNN*	*PMGI*	*SDNet*	*U2Fusion*	*SwinFusion*	*CSMCA*	*EMFusion*	*DDcGAN*	*IFCNN*	*PMGI*	*SDNet*	*U2Fusion*	*SwinFusion*
FIM	0.8488	*0.8557*	0.8128	0.8266	0.8220	0.8166	0.8175	**0.8692**	*0.8718*	0.8551	0.8200	0.8542	0.8448	0.8212	0.8499	**0.8724**
Q_{abf}	0.7389	*0.7438*	0.5852	0.6634	0.1442	0.3335	0.5326	**0.7590**	0.6206	0.4854	0.2636	*0.6233*	0.3546	0.4322	0.4829	**0.6529**
SSIM	0.9469	0.9420	0.8735	*0.9477*	0.5633	0.8118	0.9086	**0.9684**	0.9264	0.8803	0.7055	**0.9564**	0.7721	0.8868	0.9133	*0.9497*
PSNR	63.2619	62.3660	61.0884	63.0737	59.7865	*63.4674*	**63.8558**	61.4250	62.4757	61.8312	59.1265	62.4618	57.8346	*62.7760*	**63.0949**	60.7366

(a) Infrared (b) Visible (c) GTF (d) DenseFuse (e) IFCNN (f) SDNet (g) U2Fusion (h) Ours

Fig. 7. Qualitative comparison of SwinFusion with five state-of-the-art methods on visible and infrared image fusion. From left to right: infrared image, visible image, and the results of GTF, DenseFuse, IFCNN SDNet, U2Fusion, and our SwinFusion.

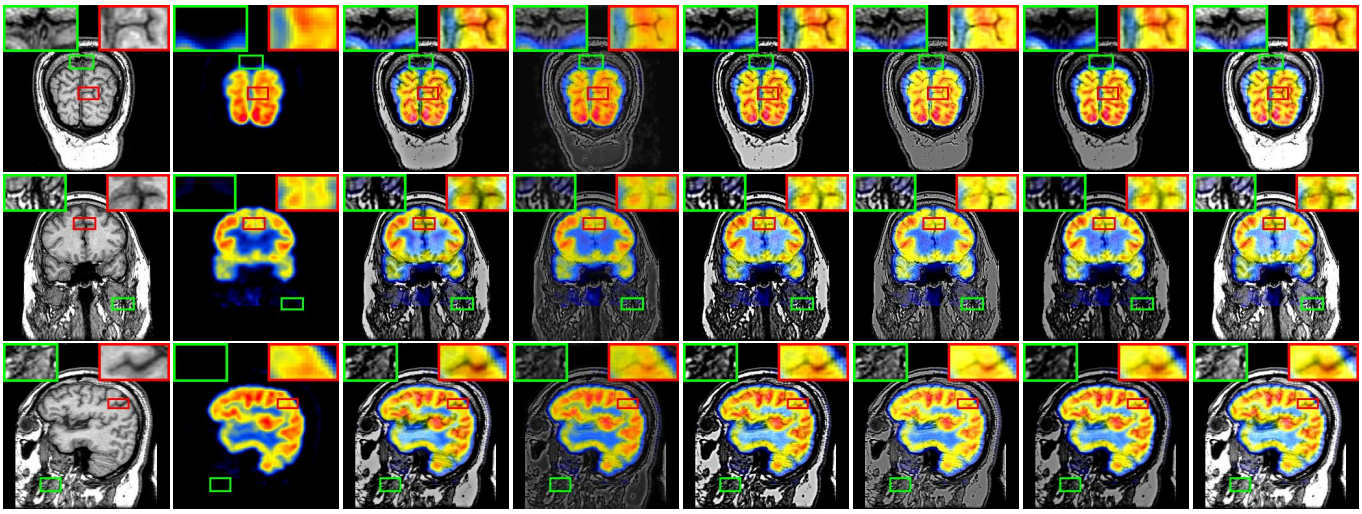
(a) Near-infrared (b) Visible (c) ANVF (d) DenseFuse (e) IFCNN (f) SDNet (g) U2Fusion (h) Ours

Fig. 8. Qualitative comparison of SwinFusion with five state-of-the-art methods on visible and near-infrared image fusion. From left to right: near-infrared image, visible image, and the results of ANVF, DenseFuse, IFCNN, SDNet, U2Fusion, and our SwinFusion.

phenomenon also occurs in medical image fusion because our fusion network focuses more on significant regions and ignores unimportant areas in source images.

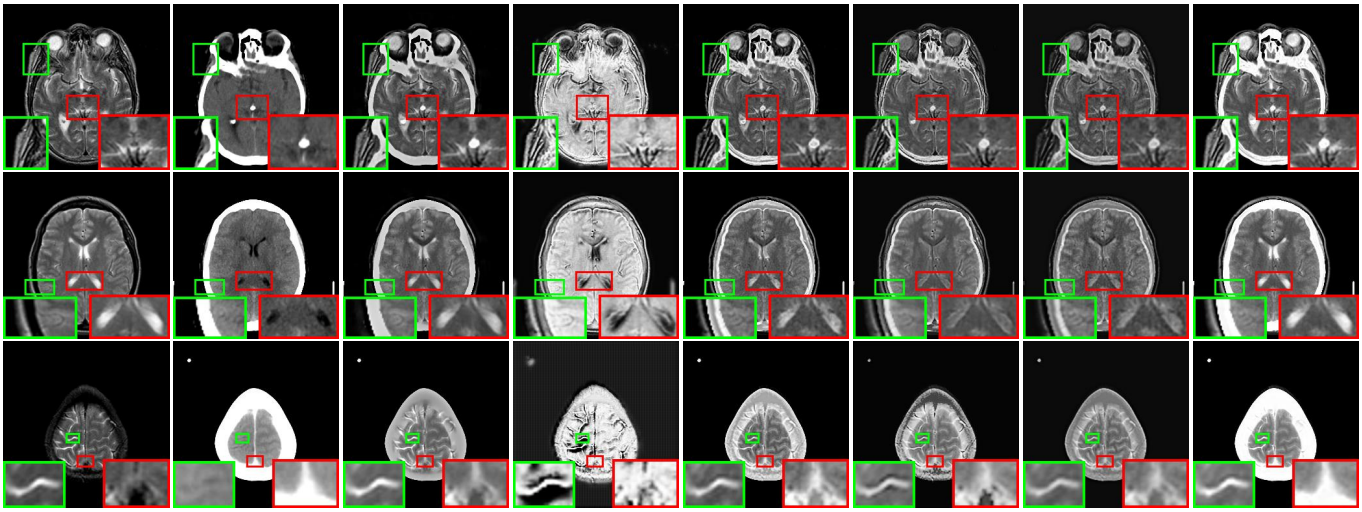
Visual Quality Comparison: We also provide some visual results in Fig. 7 - Fig. 10 to intuitively exhibit the advantages of our method in global context integration. As can be seen in Fig. 7, GTF, SDNet, and U2Fusion cannot effectively present scene information in the visible images, since the lack of global information interaction and inappropriate intensity control. In addition, DenseFuse and IFCNN can preserve some of the texture details of visible images, but still suffer from thermal radiation contamination and weaken salient targets of infrared images to varying degrees. It is worth emphasizing that our SwinFusion not only preserves the scene information of visible images but also maintains the salient objects, benefiting from effective global context perception and proper intensity control. In particular, our model is able to adaptively attend to salient regions in infrared images and the background in the visible images through the intra- and inter-modal long-range modeling and global context aggregation.

For visible and near-infrared image fusion, an excellent fusion algorithm is expected to transfer the texture details from



(a) MRI (b) PET (c) CSMCA (d) DDcGAN (e) IFCNN (f) SDNet (g) U2Fusion (h) Ours

Fig. 9. Qualitative comparison of SwinFusion with five state-of-the-art methods on PET and MRI image fusion. From left to right: MRI image, PET image, and the results of CSMCA, DDcGAN, IFCNN, SDNet, U2Fusion, and our SwinFusion.



(a) MRI (b) CT (c) CSMCA (d) DDcGAN (e) IFCNN (f) SDNet (g) U2Fusion (h) Ours

Fig. 10. Qualitative comparison of SwinFusion with five state-of-the-art methods on CT and MRI image fusion. From left to right: MRI image, CT image, and the results of CSMCA, DDcGAN, IFCNN, SDNet, U2Fusion, and our SwinFusion.

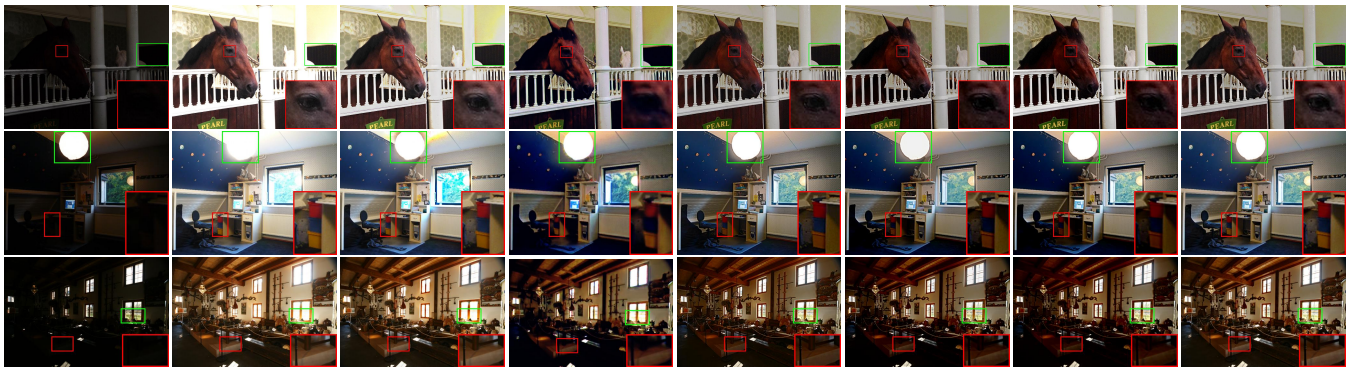
the near-infrared image into the visible image to generate the fused image. As presented in Fig. 8, ANVF, DenseFuse, and U2Fusion fail to integrate the texture details in the near-infrared images into the fused results. Only the fused images generated by IFCNN, SDNet, and SwinFusion look like sharpened visible images. Particularly, our method is superior in quantitative evaluations thanks to adequate global information aggregation, effective structure maintenance, and texture preservation.

The visual quality comparison of PET and MRI image fusion (Med (PET-MRI)) is exhibited in Fig. 9. From the results, one can find that other fusion algorithms inevitably weaken essential information in source images. More specifically, in some areas where PET images do not contain functional information, other competitors usually corrupt soft-tissue information in MRI images due to the lack of global context integration and appropriate intensity control. This issue can be observed from the green boxes in Fig. 9. In addition, as shown in red boxes, DDcGAN and SDNet fail to efficiently aggregate complementary information in source images and smooth the texture details in MRI images. It is worth noting that our fusion model can preserve the abundant details in MRI images and characterize the functional information in PET images sufficiently, owing to effective structure maintenance, global interaction aggregation, and proper intensity control.

We also provide some qualitative fused results on three typical CT and MRI image pairs in Fig. 10. In the results of other alternatives, the dense structures in CT images are weakened to different extents. Moreover, the edges in MRI images are diminished by the CSMCA, IFCNN, and U2Fusion, as shown in the first and second rows. Besides, DDcGAN fails to maintain the intensity distribution and contrast of source images. On the contrary, our SwinFusion preserves more structural (texture) information under the premise of little loss of soft-tissue details and anatomical information.

TABLE II
QUANTITATIVE COMPARISON WITH STATE-OF-THE-ARTS IN THE DIGITAL PHOTOGRAPHY IMAGE FUSION SCENARIOS. RED INDICATES THE BEST RESULT AND BLUE INDICATES THE SECOND BEST RESULT

	MEF								*MFF*							
	SPD-MEF	***MEFNet***	***MEF-GAN***	***IFCNN***	***PMGI***	***SDNet***	***U2Fusion***	***Ours***	***SFMD***	***DRPL***	***MFF-GAN***	***IFCNN***	***PMGI***	***SDNet***	***U2Fusion***	***Ours***
FMI	0.8880	0.8900	0.8741	0.8846	0.8876	*0.8915*	0.8856	**0.9015**	0.8864	**0.9016**	0.8902	0.8965	0.8852	0.8926	0.8885	*0.8979*
Q_{abf}	0.6445	0.6449	0.4667	0.6406	0.6205	*0.6754*	0.6217	**0.7581**	0.6428	0.4798	0.6694	*0.7243*	0.5784	0.6908	0.6677	**0.7365**
SSIM	0.9325	0.9418	0.8820	0.9622	0.9575	*0.9712*	0.9544	**0.9885**	0.9662	0.9884	0.9768	*0.9893*	0.9399	0.9831	0.9739	**0.9904**
PSNR	58.7611	57.8366	59.2717	*59.8078*	59.5107	59.6870	59.5709	**59.8245**	73.4801	75.2730	72.7979	*75.5575*	73.5020	73.6039	72.7259	**76.0232**



(a) Under-exposed (b) Over-exposed (c) SPD-MEF (d) MEF-GAN (e) IFCNN (f) SDNet (g) U2Fusion (h) Ours

Fig. 11. Qualitative results of multi-exposure image fusion. From left to right: under-exposed image, over-exposed image, and the results of SPD-MEF, MEF-GAN, IFCNN SDNet, U2Fusion, and our SwinFusion.

D. Results on Digital Photography Image Fusion

Quantitative Comparison: The quantitative comparisons between our method and other alternatives on digital photography image fusion scenarios are exhibited in Table II. From the results, we see that our framework ranks first in Q_{abf}, SSIM, and PSNR for both multi-exposure image fusion and multi-focus image fusion. Moreover, the proposed method achieves the best FMI for MEF, and it only lags behind MADCNN by a narrow margin in the FMI metric for MFF. The above phenomena indicate that our model can effectively integrate the complementary information and sufficiently maintain texture and structure information in source images.

Visual Quality Comparison: The qualitative comparisons of multi-exposure image fusion are shown in Fig. 11. As one can observe, other algorithms fail to maintain appropriate exposure levels since these algorithms lack the ability of global exposure awareness. More specifically, SDNet and U2Fusion are unable to light up the scene information hidden in the darkness (*e.g.*, regions highlighted in the red boxes). Although the overall exposure level of MEF-GAN is slightly better, it causes local over-exposure and blur, owing to the introduction of down-sampling in the modeling process. SPD-MEF and IFCNN introduce artifacts in some areas, such as the light in the second row. In addition, SPD-MEF loses all information in the under-exposed images, leading to severe over-exposure in the fused results. Only our SwinFsuion can effectively merge the complementary information in the source images and maintain the appropriate exposure level by global exposure perception.

We also present the results of subjective comparison for multi-focus image fusion in Fig. 12. From the results, we can note that all approaches could integrate information from the focused regions in different source images and generate an all-in-focus image. However, MFF-AGN, SDNet and U2Fusion cannot retain the best intensity distribution due to the lack of global context interaction. Our method enables adaptive focus region awareness and maintains the proper intensity distribution through global context aggregation.

In summary, extensive objective and subjective comparisons on both multi-modal image fusion and digital photography image fusion demonstrate the superiority of our SwinFusion in terms of structure maintenance, texture detail preservation, and appropriate intensity control. We summarize our strengths in the following aspects. On the one hand, we explicitly design the corresponding loss functions to achieve structure retention, texture preservation, and adaptive intensity control separately. On the other hand, the proposed attention-guided cross-domain fusion module enables intra- and inter-domain long-range dependency modeling and global context aggregation, which allows our method to model intensity distribution from a global perspective. Moreover, the transformer-based deep feature extraction module also assists our model to mine significant features and information from the global viewpoint.

E. Visualization of Global Information

As mentioned previously, our method is able to adequately exploit global information within and between domains. Abstractly, for multi-modal image fusion, our method can accurately perceive salient features (*e.g.*, the thermal targets in the infrared images and soft-tissue information in MRI images) by combining **global information** and integrate them effectively into the fused images. For digital photography image fusion, **global information** assists our model in perceiving

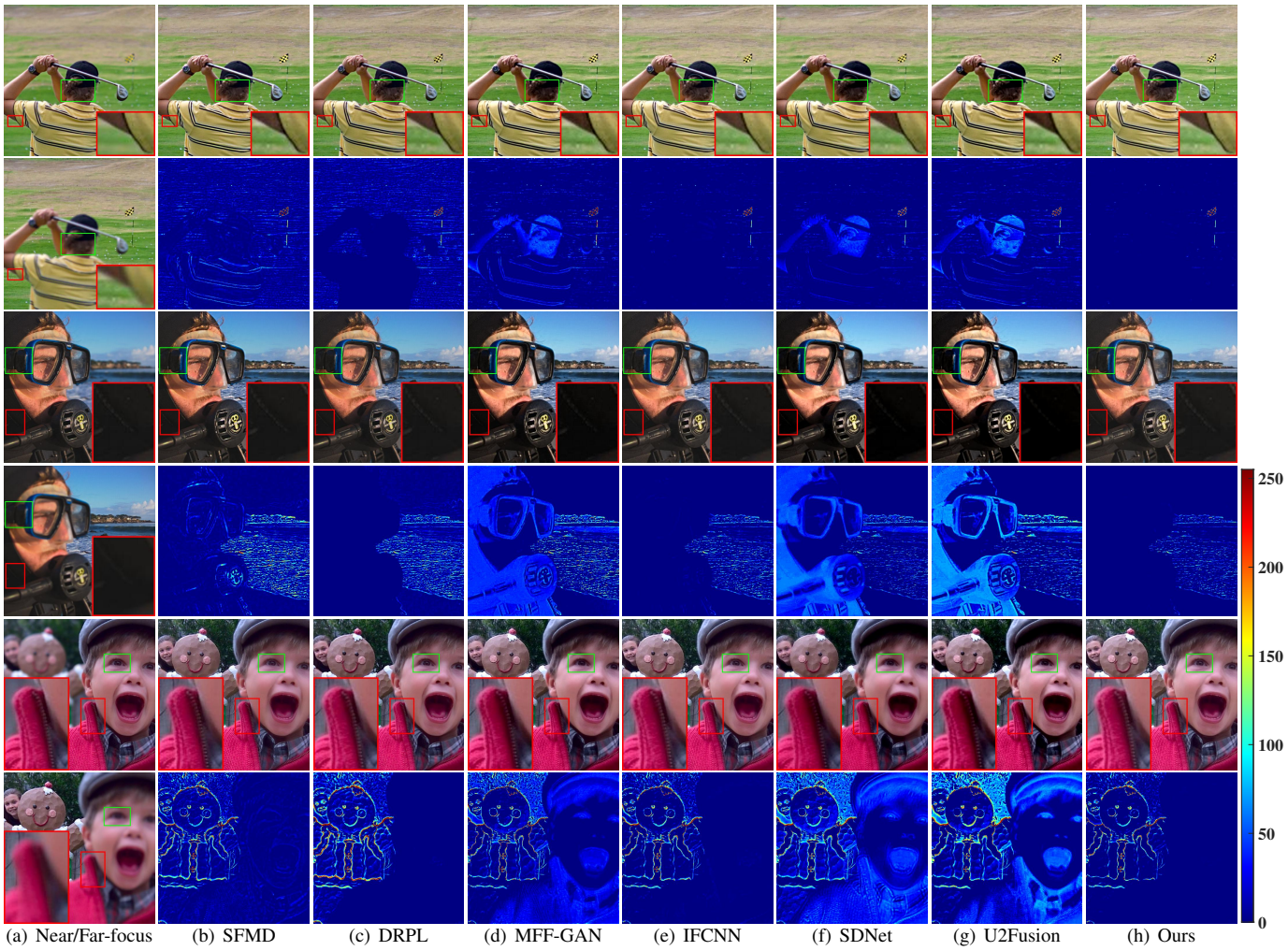


Fig. 12. Qualitative results of multi-focus image fusion. From left to right: near/far-focus image, the fused results and difference maps of SFMD, DRPL, MFF-GAN, IFCNN, SDNet, U2Fusion, and our SwinFusion. The difference maps represent the difference between the near-focus image and fused results.

the intensity distribution in source images from a global viewpoint and presenting scene information with the appropriate intensity. In order to intuitively demonstrate the role of global information, we provide a schematic diagram in Fig. 13. The second column shows the fused results with the local patches as inputs. One can note that when global information is lacking (*i.e.*, with local patches as inputs), our model fails to preserve the prominent targets in the infrared images effectively. Besides, without the ability to perceive the intensity distribution from a global perspective, our method is also unable to present scene information with a proper exposure level in the multi-exposure image fusion task. Specifically, the fused images suffer from alternating underexposure and normal exposure. On the contrary, when the whole images are used as inputs, providing enough global information for our model, our method not only effectively preserves the salient targets in the infrared images but also presents scene information with normal exposure.

F. Image Fusion for Other Vision Tasks

We investigate the positive role of image fusion for other vision tasks. Specifically, we analyze the performance of other vision missions (*e.g.*, semantic segmentation, object detection, and depth estimation) with the conditions of taking the source and fused images as inputs.

TABLE III
SEGMENTATION PERFORMANCE (MIOU) OF VISIBLE, INFRARED AND FUSED IMAGES ON THE MSRS DATASET. **RED** DENOTES THE BEST RESULT

	Person	Car	Bike	Curve	mIoU
Visible	59.94	89.03	70.00	60.69	69.92
Infrared	70.46	87.33	69.23	58.74	71.44
Ours	71.78	89.47	70.87	61.18	73.33

Visible and Infrared Image Fusion for Semantic Segmentation: The relevant experimental configuration is followed by SeAFusion [7]. The quantitative results of semantic segmentation, measured by pixel intersection-over-union (IoU), are exhibited in Table III. From the results, one can see that our fusion method could effectively facilitate the segmentation model [86] to perceive the imaging scenario by adequately integrating intra- and inter-modal complementary information as well as global context. We also provide some visual examples in Fig. 14 to intuitively reveal the positive effect of fused results

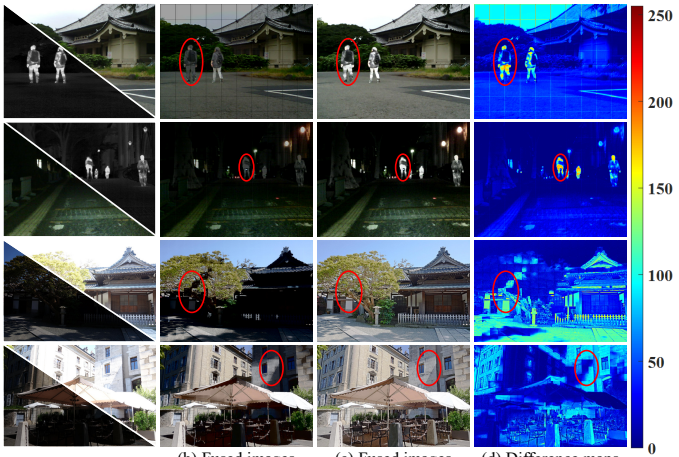


Fig. 13. Visualization of global information. The first two rows represent the visible and infrared image fusion scenario, and the following two rows show the multi-exposure image fusion scenario.

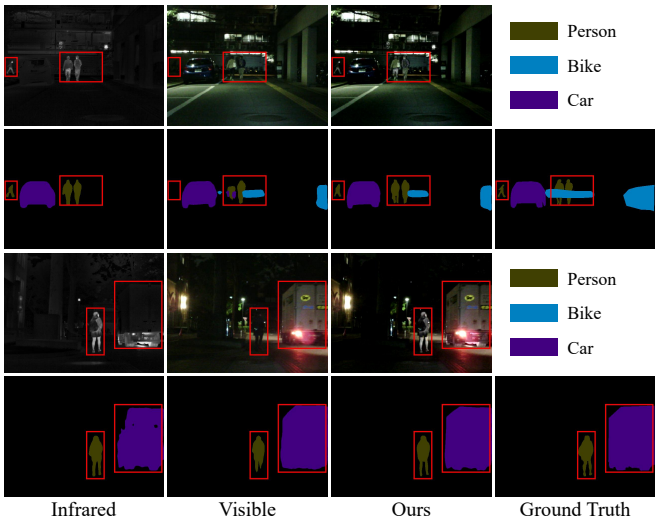


Fig. 14. Visual results of semantic segmentation. The first and third rows represent the source image and fused images. The second and fourth rows exhibit the corresponding segmentation results and ground truth.

for semantic segmentation. One can note that the infrared image can provide enough information about pedestrians and cars, but cannot give information about the bike for the segmentation model in the first scene. In contrast, the segmentation model can segment the car and bikes, but fails to split people from the visible image completely. Specifically, the segmentation model can segment the pedestrians, cars, and bikes from the fused image, which integrates the advantages of both infrared and visible images. Moreover, neither the visible nor infrared images provide enough information for the segmentation model to completely split both person and car in the second scenario. The segmentation network only perceives adequate scene information from the fused image to fully segment the car and pedestrian.

Visible and Infrared Image Fusion for Object Detection: We also investigate the role of visible and infrared image fusion in object detection. A state-of-the-art detection network,*i.e.*,

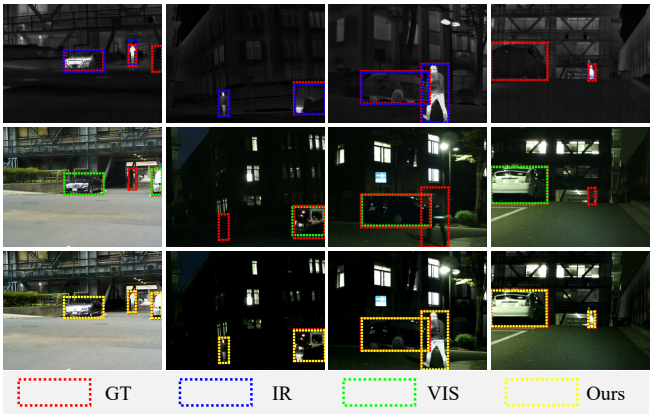


Fig. 15. Visual results of object detection. From top to bottom are the detection results of infrared images, visible images, and fused images generated by SwinFusion, respectively.

YOLOv5 [87] is employed to measure object detection performance on the source images and fused images. The test set is collected and labeled by GAN-FM [88]. We exhibit the mean average precision (mAP) of object detection in Table IV, where AP@0.5, AP@0.7 and AP@0.9 denote the AP values at IoU thresholds of 0.5, 0.7 and 0.9, respectively, and mAP@[0.5:0.95] stands for the average of all AP values at different IoU thresholds (from 0.5 to 0.95 in steps of 0.05).

As shown in Table IV, we can observe that visible and infrared images provide only object-specific information for the detector. Thus, the detection model achieves better car detection performance on visible images, but has superior pedestrian detection performance on infrared images. Such complementary characteristics offer the potential for the detector to achieve better performance on fused images. In fact, the detection network shows a more balanced performance on fused images. Moreover, the fused images can provide a more comprehensive description for cars by combining complementary information in source images. Thereby, the detector could achieve better car detection performance. Although the performance of pedestrian detection on fused images is inferior to that on infrared images, it is justifiable. The infrared image merely collects the thermal radiation information from the salient object but ignores the surrounding environment, resulting in a higher contrast for prominent targets such as person, which facilitates the detector to detect pedestrians. Some visualized examples are presented in Fig. 15.

Multi-focus Image Fusion for Depth Estimation: The impacts of multi-focus images and fused results on depth estimation are shown in Fig. 16. We employ AdaBins [89] to estimate the high quality dense depth map from a single RGB input image. From the visual results, we can find that AdaBins only successfully estimates the dense depth map of objects in the focused regions from the multi-focus images, while objects in non-focused areas are treated equally, *i.e.*, the correct depth map cannot be estimated. Moreover, our method effectively integrates the scene information in the focused regions of the source images into a single all-in-focus image. Thus, AdaBins is able to successfully estimate the dense depth maps of all objects from the fused images.

TABLE IV
OBJECT DETECTION PERFORMANCE (MAP) OF VISIBLE, INFRARED AND FUSED IMAGES ON THE MSRS DATASET. **RED** INDICATES THE BEST RESULT

	AP@0.5			AP@0.7			AP@0.9			mAP@[0.5:0.95]		
	Person	Car	Average	Person	Car	Average	Person	Car	Average	Person	Car	Average
Infrared	0.962	0.818	0.890	0.888	0.737	0.813	0.293	0.237	0.265	0.706	0.625	0.665
Visible	0.737	0.948	0.843	0.457	0.873	0.665	0.047	0.437	0.242	0.383	0.744	0.563
Fusion	0.939	0.946	0.932	0.829	0.880	0.854	0.158	0.550	0.354	0.637	0.762	0.699

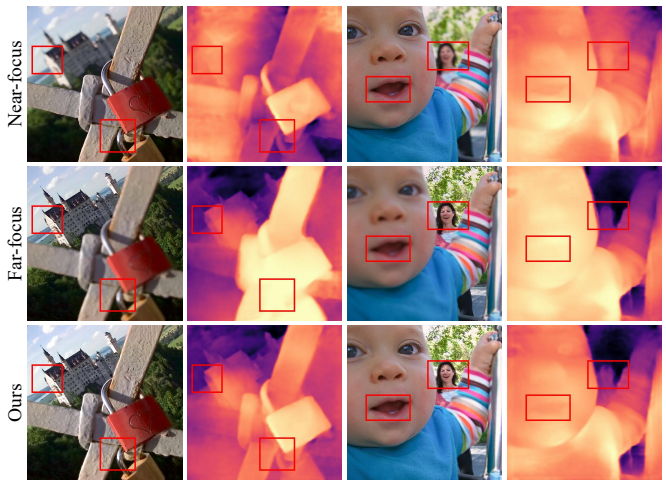


Fig. 16. Visual results of depth estimation. The first and third columns are the source images and fused results. The second and fourth columns are the corresponding dense depth maps.

G. Ablation Study

The performance of our SwinFusion relies on the elaborate network architecture and loss functions. On the one hand, the joint CNN-Transformer architecture effectively mines the local information and global interaction from the source images. Especially, the Transformer-based deep feature extraction (DE) fully extracts the global context in the shallow features. Moreover, the attention-guided cross-domain fusion module (ACFM) adequately integrates complementary information as well as intra- and inter-domain long-range dependencies, which allows our network to perceive apparent intensity from a global perspective. On the other hand, the designed SSIM loss, texture loss, and intensity loss drive our model to achieve effective structure maintenance, texture detail preservation, and appropriate intensity control. In this section, we perform a series of ablation studies to verify the effectiveness of specific designs. The visual results of the ablation experiments on the multi-modal image fusion (*e.g.*, VIF) and digital photography image fusion (*e.g.*, MEF) are presented in Fig. 17.

Analysis of Deep Feature Extraction (DE): The Transformer-based deep feature extraction could exploit the global context in the shallow features to provide an appropriate intensity perception for the fusion model. As shown in Fig. 17 (c), the fused results fail to present a suitable apparent intensity after removing the deep feature extraction. Specifically, the fusion network is unable to perceive significant and complementary information in the source images for visible and infrared image fusion.

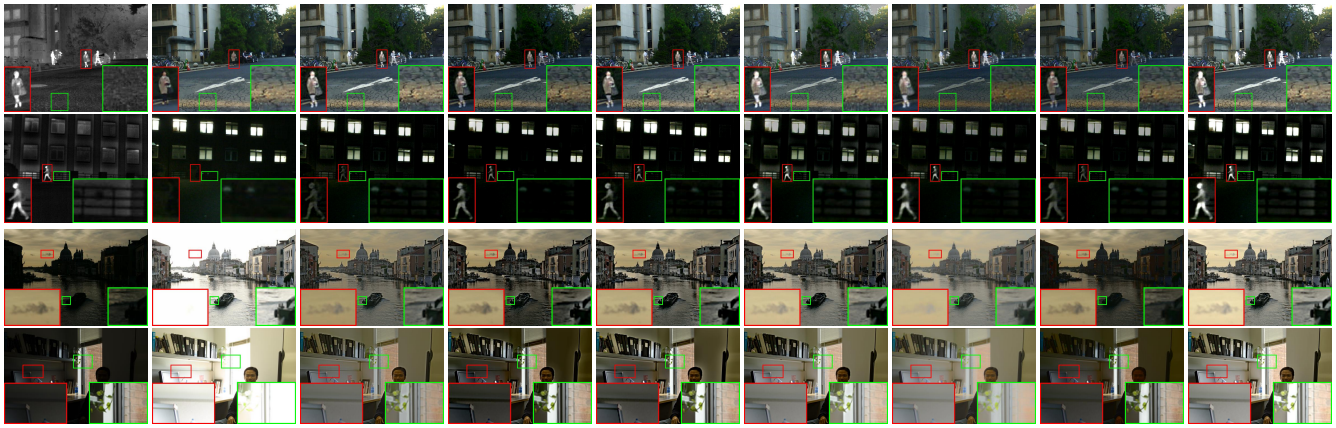
Analysis of Attention-guided Cross-domain Fusion Module (ACFM): The attention-guided cross-domain fusion module consists of the intra-domain fusion unit based on self-attention and the inter-domain fusion unit based on cross-attention, which can fully aggregate long-range dependencies and global interaction within the same domain and across domains. From Fig. 17 (d), we can find that the fusion model cannot effectively control the apparent intensity of the fused images after removing the ACFM. This phenomenon is particularly evident for the multi-exposure image fusion task, *i.e.*, the fused results fail to present an appropriate exposure level.

Analysis of Inter-domain Fusion Unit (Inter): The inter-domain fusion unit, an essential component of the attention-guided cross-domain fusion module, is able to fully exploit and leverage the long-range dependencies across domains to achieve effective information integration. As illustrated in Fig. 17 (e), the visual performance of the network without the inter-domain fusion unit on the VIF task is similar to that of the network without the ACFM. This phenomenon indicates that significant targets and structures in the VIF task are perceived by integrating global information across domains. In addition, removing only the inter-domain fusion unit improves the exposure situation of the fused images compared to removing the whole ACFM. However, the fused images generated by the network without the inter-domain fusion unit still fail to present the scene information with a suitable exposure level.

Analysis of Structure Maintenance ($\mathcal{L}_{ssim}$): We introduce SSIM loss ($\mathcal{L}_{ssim}$) to constrain the fusion network to maintain the structural information in the source images. In addition, SSIM loss could restrain the brightness of the fusion results to some extent. As presented in Fig. 17 (f), the network without the constraint of SSIM loss cannot maintain the optimal structure and intensity information. In particular, the salient targets in the infrared images are slightly weakened for the VIF scenario.

Analysis of Texture Preservation ($\mathcal{L}_{text}$): For better characterizing the imaging scenes, the abundant textures of source images are expected to be preserved in the fused results as much as possible. Thus, we design texture loss to retain more texture details. From Fig. 17 (g), one can note that the fusion model trained without texture loss fails to generate fused images with rich textures. This issue can be observed in cracks on the ground, fences on the roadside, clouds in the sky, and textures on the wall.

Analysis of Appropriate Intensity Control ($\mathcal{L}_{int}$): We also devise intensity loss to constrain the fusion network to present fused results with proper intensity. The fused images without the constraint of intensity loss are exhibited in Fig. 17 (h).



(a) Source image1 (b) Source image2 (c) w/o DE (d) w/o ACFM (e) w/o Inter (f) w/o $\mathcal{L}_{ssim}$ (g) w/o $\mathcal{L}_{text}$ (h) w/o $\mathcal{L}_{int}$ (i) Ours

Fig. 17. Visual results of the ablation experiment on the visible and infrared image fusion and multi-exposure image fusion scenarios.

TABLE V
QUANTITATIVE EVALUATION RESULTS OF ABLATION STUDIES ON THE VIF AND MEF TASKS. **RED** DENOTES THE BEST RESULT

		w/o DE	w/o ACFM	w/o Inter	w/o $\mathcal{L}_{ssim}$	w/o $\mathcal{L}_{text}$	w/o $\mathcal{L}_{int}$	Ours
VIF	FMI	0.9036	0.9073	0.9019	0.9290	0.9173	0.9142	**0.9308**
	Q_{abf}	0.3355	0.4592	0.5612	0.2374	0.3809	0.4742	**0.6428**
	SSIM	0.8906	0.9403	0.9404	0.8833	0.9292	0.9404	**0.96976**
	PSNR	66.0046	66.4919	64.1470	66.8560	**67.0619**	66.7691	64.4063
MEF	FMI	0.8874	0.8991	0.9011	0.8971	0.8836	0.9010	**0.9015**
	Q_{abf}	0.5510	0.7173	0.7252	0.7109	0.3012	0.3737	**0.7581**
	SSIM	0.9305	0.9763	0.9845	0.9695	0.8587	0.8231	**0.9885**
	PSNR	58.9233	59.2280	58.4322	59.6521	59.5344	57.5740	**59.8245**

From the results, we can find that the fusion network is unable to generate the fused results with suitable apparent intensity after removing intensity loss. Specifically, the fusion model weakens the significant objects for the VIF task and fails to perceive the normal exposure level for the MEF scenario.

It is worth emphasizing that our SwinFusion can implement effective structure maintenance, texture preservation, and proper intensity control with the constraint of SSIM loss, texture loss, and intensity loss. In particular, our fusion model could achieve global intensity perception by sufficiently integrating intra- and inter-domain long-range dependencies and global interactions.

The quantitative results of ablation studies are presented in Table V. As we can see from the results, removing any of the components will degrade the fusion performance to a greater or lesser extent. The PSNR is improved for the visible and infrared image fusion, which is caused by the fusion model failing to perceive significant information in the source images. Especially, although the degradation of visual results is not severe after removing SSIM loss, there is a distinct degradation of the quantitative evaluation.

H. Computational Complexity Analysis

As shown in Table VI, a complexity evaluation is introduced to evaluate the operational efficiency of different methods from three perspectives, *i.e.*, training parameters (size), floating-point operations per second (FLOPs), and runtime. The first image of the test set in each fusion scenario is utilized to calculate

TABLE VI
COMPUTATIONAL EFFICIENCY COMPARISON WITH STATE-OF-THE-ART IMAGE FUSION METHODS. **RED** AND BLUE DENOTE THE BEST AND THE SECOND BEST RESULTS, RESPECTIVELY

	Method	**GTF**	**DenseFuse**	**FusionGAN**	**IFCNN**	**PMGI**	**SDNet**	**U2Fusion**	**Ours**
VIF	**Size(M)**	–	0.0742	0.9256	0.0836	**0.0420**	*0.0671*	0.6592	0.9737
	FLOPs(G)	–	54.1512	576.2035	*39.932*	**25.6380**	41.3107	405.1745	292.5376
	Time(s)	3.8878	0.1767	0.0589	*0.0130*	0.0520	**0.0125**	0.1426	1.4668
	Method	**ANVF**	**DenseFuse**	**GANMcC**	**IFCNN**	**PMGI**	**SDNet**	**U2Fusion**	**Ours**
VIS-NIR	**Size(M)**	–	0.0742	1.8641	0.0836	**0.0420**	*0.0671*	0.6592	0.9737
	FLOPs(G)	–	122.5621	2590.2397	*90.6457*	**58.0273**	93.4999	917.0432	649.1954
	Time(s)	79.5474	0.2956	0.7376	**0.0276**	0.1471	*0.077*	0.6019	3.3839
	Method	**CSMCA**	**EMFusion**	**DDcGAN**	**IFCNN**	**PMGI**	**SDNet**	**U2Fusion**	**Ours**
Med	**Size(M)**	–	0.1485	1.0977	0.0836	**0.0420**	*0.0671*	0.6592	0.9737
	FLOPs(G)	–	19.4851	211.6094	*8.5188*	**5.4694**	8.8130	86.4383	64.4755
	Time(s)	27.3876	0.1102	0.7676	**0.0265**	0.1265	*0.1062*	0.4393	0.2748
	Method	**SPD-MEF**	**MEFNet**	**MEF-GAN**	**IFCNN**	**PMGI**	**SDNet**	**U2Fusion**	**Ours**
MEF	**Size(M)**	–	**0.0265**	0.4876	0.0836	*0.0420*	0.0671	0.6592	0.9737
	FLOPs(G)	–	**1.2929**	178.3240	24.8036	*15.9249*	25.6600	251.6729	181.1708
	Time(s)	16.4619	0.1449	1.5527	**0.0204**	0.1532	*0.1011*	0.8228	2.0047
	Method	**SFMD**	**DRPL**	**MFF-GAN**	**IFCNN**	**PMGI**	**SDNet**	**U2Fusion**	**Ours**
MFF	**Size(M)**	–	1.0700	**0.0400**	0.0836	*0.0420*	0.0671	0.6592	0.9737
	FLOPs(G)	–	289.3329	*21.7015*	*35.1485*	22.5668	36.3620	356.6381	257.9020
	Time(s)	0.2729	0.0994	0.1167	**0.0315**	0.1236	*0.0949*	0.5412	1.3545

the FLOPS of methods. From the results, we can observe that the deep learning-based methods have a significant advantage in runtime compared to traditional methods, benefiting from GPU acceleration. Among the general image fusion methods, PMGI, IFCNN, and SDNet have lower training parameters, FLOPs, and average running time. Moreover, MEFNet and MFF-GAN have the lowest training parameters and FLOPs in the MEF and MFF tasks, respectively. It is worth pointing out that our method has comparable operational efficiency with the dominant image fusion algorithms, although it requires computing pixel-to-pixel correlations (*i.e.*, attention) within a window and includes several transformer-based components.

V. CONCLUSION

In this paper, we have proposed a general image fusion method based on cross-domain long-range learning and Swin Transformer, called SwinFusion, which could handle multi-modal image fusion and digital photography image fusion in

a unified framework. The proposed SwinFusion unifies multi-modal image fusion and digital photography image fusion as structure maintenance, texture detail preservation, and appropriate intensity control. Then, we have devised a unified loss function consisting of SSIM loss, texture loss, and intensity loss to constrain the network to fulfill the corresponding functions. Moreover, a self-attention-based intra-domain fusion unit and a cross-attention-based inter-domain fusion unit have been developed to adequately integrate the long-range dependencies and global interactions within the same domain and across domains. Based on the elaborate network architecture and loss functions, the proposed method is able to maintain the structural information and preserve abundant texture details in source images under both the multi-modal image fusion and digital photography image fusion scenarios. In addition, our model presents the appropriate apparent intensity for the fused image from a global perspective. Extensive experiments have been conducted to verify the superiority of SwinFusion compared to state-of-the-art alternatives. Besides, the expanded experiments on semantic segmentation, object detection, and depth estimation demonstrate the potential of image fusion for other computer vision tasks.

REFERENCES

[1] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, "Image fusion meets deep learning: A survey and perspective," *Information Fusion*, vol. 76, pp. 323–336, 2021.

[2] J. Ma, L. Tang, M. Xu, H. Zhang, and G. Xiao, "Stdfusionnet: An infrared and visible image fusion network based on salient target detection," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, p. 5009513, 2021.

[3] M. Awad, A. Elliethy, and H. A. Aly, "Adaptive near-infrared and visible fusion for fast image enhancement," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 408–418, 2019.

[4] H. Xu and J. Ma, "Emfusion: An unsupervised enhanced medical image fusion network," *Information Fusion*, vol. 76, pp. 177–186, 2021.

[5] Y. Liu, X. Chen, J. Cheng, and H. Peng, "A medical image fusion method based on convolutional neural networks," in *Proceedings of the International Conference on Information Fusion*, 2017, pp. 1–7.

[6] J. Ma, Z. Le, X. Tian, and J. Jiang, "Smfuse: Multi-focus image fusion via self-supervised mask-optimization," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 309–320, 2021.

[7] L. Tang, J. Yuan, and J. Ma, "Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network," *Information Fusion*, vol. 82, pp. 28–42, 2022.

[8] G. Bhatnagar and Q. J. Wu, "A fractal dimension based framework for night vision fusion," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 1, pp. 220–227, 2018.

[9] Y. Zhang and Q. Ji, "Active and dynamic information fusion for facial expression understanding from image sequences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 699–714, 2005.

[10] 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, no. 1, pp. 502–518, 2022.

[11] H. Xu, X. Wang, and J. Ma, "Drf: Disentangled representation for visible and infrared image fusion," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, p. 5006713, 2021.

[12] Q. Han and C. Jung, "Deep selective fusion of visible and near-infrared images using unsupervised u-net," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[13] K. Ram Prabhakar, V. Sai Srikar, and R. Venkatesh Babu, "Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4714–4722.

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

[15] Y. Liu, S. Liu, and Z. Wang, "A general framework for image fusion based on multi-scale transform and sparse representation," *Information Fusion*, vol. 24, pp. 147–164, 2015.

[16] H. Zhang, H. Xu, Y. Xiao, X. Guo, and J. Ma, "Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity." in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 12 797–12 804.

[17] 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.

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

[19] W. Zhao, H. Lu, and D. Wang, "Multisensor image fusion and enhancement in spectral total variation domain," *IEEE Transactions on Multimedia*, vol. 20, no. 4, pp. 866–879, 2017.

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

[21] F. Zhao, W. Zhao, L. Yao, and Y. Liu, "Self-supervised feature adaption for infrared and visible image fusion," *Information Fusion*, vol. 76, pp. 189–203, 2021.

[22] H. Li and X.-J. Wu, "Densefuse: A fusion approach to infrared and visible images," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2614–2623, 2019.

[23] H. Xu, H. Zhang, and J. Ma, "Classification saliency-based rule for visible and infrared image fusion," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 824–836, 2021.

[24] 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.

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

[26] J. Huang, Z. Le, Y. Ma, F. Fan, H. Zhang, and L. Yang, "Mgmdcgan: Medical image fusion using multi-generator multi-discriminator conditional generative adversarial network," *IEEE Access*, vol. 8, pp. 55 145–55 157, 2020.

[27] J. Chen, X. Li, L. Luo, X. Mei, and J. Ma, "Infrared and visible image fusion based on target-enhanced multiscale transform decomposition," *Information Sciences*, vol. 508, pp. 64–78, 2020.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 1–11.

[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proceedings of the International Conference on Learning Representations*, 2020, pp. 1–12.

[30] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 213–229.

[31] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6881–6890.

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

[33] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 1833–1844.

[34] 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 preprint arXiv:2112.01030*, 2021.

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

[36] D. Rao, X.-J. Wu, and T. Xu, "Tgfuse: An infrared and visible image fusion approach based on transformer and generative adversarial network," *arXiv preprint arXiv:2201.10147*, 2022.

[37] Y. Fu, T. Xu, X. Wu, and J. Kittler, "Ppt fusion: Pyramid patch transformerfor a case study in image fusion," *arXiv preprint arXiv:2107.13967*, 2021.

[38] 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," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[39] Y. Liu, X. Chen, R. K. Ward, and Z. J. Wang, "Medical image fusion via convolutional sparsity based morphological component analysis," *IEEE Signal Processing Letters*, vol. 26, no. 3, pp. 485–489, 2019.

[40] Y. Liu, S. Liu, and Z. Wang, "Multi-focus image fusion with dense sift," *Information Fusion*, vol. 23, pp. 139–155, 2015.

[41] K. Ma, H. Li, H. Yong, Z. Wang, D. Meng, and L. Zhang, "Robust multi-exposure image fusion: a structural patch decomposition approach," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2519–2532, 2017.

[42] H. Li, L. Li, and J. Zhang, "Multi-focus image fusion based on sparse feature matrix decomposition and morphological filtering," *Optics Communications*, vol. 342, pp. 1–11, 2015.

[43] H. Li, X.-J. Wu, and J. Kittler, "Infrared and visible image fusion using a deep learning framework," in *Proceedings of the International Conference on Pattern Recognition*, 2018, pp. 2705–2710.

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

[45] H. Ma, Q. Liao, J. Zhang, S. Liu, and J.-H. Xue, "An α-matte boundary defocus model-based cascaded network for multi-focus image fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 8668–8679, 2020.

[46] J. Li, X. Guo, G. Lu, B. Zhang, Y. Xu, F. Wu, and D. Zhang, "Drpl: Deep regression pair learning for multi-focus image fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 4816–4831, 2020.

[47] F. Zhao, W. Zhao, H. Lu, Y. Liu, L. Yao, and Y. Liu, "Depth-distilled multi-focus image fusion," *IEEE Transactions on Multimedia*, 2021.

[48] W. Zhao, B. Zheng, Q. Lin, and H. Lu, "Enhancing diversity of defocus blur detectors via cross-ensemble network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8905–8913.

[49] W. Zhao, X. Hou, Y. He, and H. Lu, "Defocus blur detection via boosting diversity of deep ensemble networks," *IEEE Transactions on Image Processing*, vol. 30, pp. 5426–5438, 2021.

[50] D. Han, L. Li, X. Guo, and J. Ma, "Multi-exposure image fusion via deep perceptual enhancement," *Information Fusion*, vol. 79, pp. 248–262, 2022.

[51] Y. Long, H. Jia, Y. Zhong, Y. Jiang, and Y. Jia, "Rxdnfuse: A aggregated residual dense network for infrared and visible image fusion," *Information Fusion*, vol. 69, pp. 128–141, 2021.

[52] H. Li, X.-J. Wu, and T. Durrani, "Nestfuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 12, pp. 9645–9656, 2020.

[53] H. Li, X.-J. Wu, and J. Kittler, "Rfn-nest: An end-to-end residual fusion network for infrared and visible images," *Information Fusion*, vol. 73, pp. 720–86, 2021.

[54] L. Jian, X. Yang, Z. Liu, G. Jeon, M. Gao, and D. Chisholm, "Sedrfuse: A symmetric encoder–decoder with residual block network for infrared and visible image fusion," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, p. 20151911, 2021.

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

[56] H. Zhang, Z. Le, Z. Shao, H. Xu, and J. Ma, "Mff-gan: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion," *Information Fusion*, vol. 66, pp. 40–53, 2021.

[57] J. Ma, W. Yu, C. Chen, P. Liang, X. Guo, and J. Jiang, "Pan-gan: An unsupervised pan-sharpening method for remote sensing image fusion," *Information Fusion*, vol. 62, pp. 110–120, 2020.

[58] J. Li, H. Huo, C. Li, R. Wang, C. Sui, and Z. Liu, "Multigrained attention network for infrared and visible image fusion," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, p. 5002412, 2021.

[59] J. Li, H. Huo, C. Li, R. Wang, and Q. Feng, "Attentionfgan: Infrared and visible image fusion using attention-based generative adversarial networks," *IEEE Transactions on Multimedia*, vol. 23, pp. 1383–1396, 2020.

[60] H. Zhang and J. Ma, "Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion," *International Journal of Computer Vision*, vol. 129, no. 10, pp. 2761–2785, 2021.

[61] F. Zhao and W. Zhao, "Learning specific and general realm feature representations for image fusion," *IEEE Transactions on Multimedia*, vol. 23, pp. 2745–2756, 2020.

[62] H. Xu, J. Ma, Z. Le, J. Jiang, and X. Guo, "Fusiondn: A unified densely connected network for image fusion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 12 484–12 491.

[63] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," in *Proceedings of the IEEE Conference on Computer Cision and Pattern Recognition*, 2021, pp. 16 519–16 529.

[64] M. Chen, H. Peng, J. Fu, and H. Ling, "Autoformer: Searching transformers for visual recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 12 270–12 280.

[65] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," in *Proceedings of the International Conference on Learning Representations*, 2020, pp. 1–11.

[66] Z. Sun, S. Cao, Y. Yang, and K. M. Kitani, "Rethinking transformer-based set prediction for object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 3611–3620.

[67] P. Sun, J. Cao, Y. Jiang, R. Zhang, E. Xie, Z. Yuan, C. Wang, and P. Luo, "Transtrack: Multiple object tracking with transformer," *arXiv preprint arXiv:2012.15460*, 2020.

[68] X. Chen, B. Yan, J. Zhu, D. Wang, X. Yang, and H. Lu, "Transformer tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8126–8135.

[69] L. Lin, H. Fan, Y. Xu, and H. Ling, "Swintrack: A simple and strong baseline for transformer tracking," *arXiv preprint arXiv:2112.00995*, 2021.

[70] H. Wang, Y. Zhu, H. Adam, A. Yuille, and L.-C. Chen, "Max-deeplab: End-to-end panoptic segmentation with mask transformers," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5463–5474.

[71] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, "Learning texture transformer network for image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5791–5800.

[72] H. Zhao and R. Nie, "Dndt: Infrared and visible image fusion via densenet and dual-transformer," in *Proceedings of the International Conference on Information Technology and Biomedical Engineering*, 2021, pp. 71–75.

[73] J. Li, J. Zhu, C. Li, X. Chen, and B. Yang, "Cgtf: Convolution-guided transformer for infrared and visible image fusion," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, p. 5012314, 2022.

[74] T. Xiao, P. Dollar, M. Singh, E. Mintun, T. Darrell, and R. Girshick, "Early convolutions help transformers see better," in *Advances in Neural Information Processing Systems*, 2021, pp. 30 392–30 400.

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

[76] L. Tang, J. Yuan, H. Zhang, X. Jiang, and J. Ma, "Piafusion: A progressive infrared and visible image fusion network based on illumination aware," *Information Fusion*, vol. 83, pp. 79–92, 2022.

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

[78] M. Brown and S. Süsstrunk, "Multi-spectral sift for scene category recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 177–184.

[79] J. Cai, S. Gu, and L. Zhang, "Learning a deep single image contrast enhancer from multi-exposure images," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 2049–2062, 2018.

[80] X. Zhang, "Benchmarking and comparing multi-exposure image fusion algorithms," *Information Fusion*, vol. 74, pp. 111–131, 2021.

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

[82] 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. 70, p. 5005014, 2021.

[83] K. Ma, Z. Duanmu, H. Zhu, Y. Fang, and Z. Wang, "Deep guided learning for fast multi-exposure image fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 2808–2819, 2020.

[84] M. B. A. Haghighat, A. Aghagolzadeh, and H. Seyedarabi, "A non-reference image fusion metric based on mutual information of image features," *Computers & Electrical Engineering*, vol. 37, no. 5, pp. 744–756, 2011.

[85] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, pp. 8026–8037.

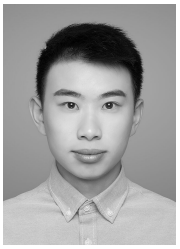
[86] C. Peng, T. Tian, C. Chen, X. Guo, and J. Ma, "Bilateral attention decoder: A lightweight decoder for real-time semantic segmentation," *Neural Networks*, vol. 137, pp. 188–199, 2021.

[87] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[88] 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.

[89] S. F. Bhat, I. Alhashim, and P. Wonka, "Adabins: Depth estimation using adaptive bins," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4009–4018.

Jiayi Ma (Senior Member, IEEE) received the B.S. degree in information and computing science and the Ph.D. degree in control science and engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2008 and 2014, respectively. He is currently a Professor with the Electronic Information School, Wuhan University. He has authored or co-authored more than 200 refereed journal and conference papers, including *IEEE TPAMI/TIP*, *IJCV*, *CVPR*, *ICCV*, *ECCV*, *etc*. His research interests include computer vision, machine learning, and robotics. Dr. Ma has been identified in the 2019-2021 Highly Cited Researcher lists from the Web of Science Group. He is an Area Editor of *Information Fusion* and an Associate Editor of *Neurocomputing*.



Linfeng Tang received the B.E. degree from the School of Computer Science and Engineering, Central South University, Changsha, China, in 2020. He is currently a Master student with the Electronic Information School, Wuhan University. His research interests include computer vision, machine learning, and pattern recognition.

Fan Fan received the B.S. degree of Communication Engineering, and the Ph.D. degree of Electronic Circuit and System, both from the Huazhong University of Science and Technology, Wuhan, China, in 2009, and 2015, respectively. He is currently an Associate Professor in the Electronic Information School, Wuhan University, China. His current research interests include infrared thermal imaging, machine learning and computer vision.

Jun Huang received the B.S. and Ph.D. degree from the Department of Electronic and Information Engineering, Huazhong University of Science and Technology, Wuhan, China, in 2008 and 2014 respectively. He is currently an Associate Professor at Electronic Information School in Wuhan University, China. His main research interest is infrared image and infrared spectrum processing.

Xiaoguang Mei received the B.S. degree in communication engineering from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2007, the M.S. degree in communications and information systems from Central China Normal University, Wuhan, in 2011, and the Ph.D. degree in circuits and systems from the HUST, in 2016. From 2010 to 2012, he was a Software Engineer with the 722 Research Institute, China Shipbuilding Industry Corporation, Wuhan. He is currently an Associate Professor with Wuhan University. His research interests include hyperspectral image processing, image fusion and machine learning.

Yong Ma graduated from the Department of Automatic Control, Beijing Institute of Technology, Beijing, China, in 1997. He received the Ph.D. degree from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2003. Between 2004 and 2006, he was a Lecturer at the University of the West of England, Bristol, U.K. Between 2006 and 2014, he was with the Wuhan National Laboratory for Optoelectronics, HUST, Wuhan, where he was a Professor of electronics. He is now a Professor with the Electronic Information School, Wuhan University. His general field of research is in signal and systems. His research interests include infrared image processing, pattern recognition, interface circuits to sensors and actuators.