CHITNet: A Complementary to Harmonious Information Transfer Network for Infrared and Visible Image Fusion

Yafei Zhang, Keying Du, Huafeng Li, Zhengtao Yu, Yu Liu

arXiv:2309.06118v5 [cs.CV] 26 Oct 2023

Abstract—Current infrared and visible image fusion (IVIF) methods go to great lengths to excavate complementary features and design complex fusion strategies, which is extremely challenging. To this end, we rethink the IVIF outside the box, proposing a complementary to harmonious information transfer network (CHITNet). It reasonably transfers complementary information into harmonious one, which integrates both the shared and complementary features from two modalities. Specifically, to skillfully sidestep aggregating complementary information in IVIF, we design a mutual information transfer (MIT) module to mutually represent features from two modalities, roughly transferring complementary information into harmonious one. Then, a harmonious information acquisition supervised by source image (HIASSI) module is devised to further ensure the complementary to harmonious information transfer after MIT. Meanwhile, we also propose a structure information preservation (SIP) module to guarantee that the edge structure information of the source images can be transferred to the fusion results. Moreover, a mutual promotion training paradigm with interaction loss is adopted to facilitate better collaboration among MIT, HIASSI and SIP. In this way, the proposed method is able to generate fused images with higher qualities. Extensive experimental results demonstrate the superiority of CHITNet over state-of-the-art algorithms in terms of visual quality and quantitative evaluations.

Index Terms—Image fusion; Complementary to harmonious information transfer; Mutual representation learning; Mutual promotion training paradigm.

I. INTRODUCTION

Visible light imaging sensors can clearly image objects visible to the human eye, whose imaging results have the advantages of high spatial resolution, rich texture details, and more in line with human visual perception. However, such sensors fail to clearly image objects in severe weather scenes such as fog and haze. Infrared sensors image through measuring the temperature difference between target and background, able to effectively avoid the influence of harsh conditions on target imaging, yet still cannot clearly image the background information with a relatively constant temperature [1]. In order to integrate the advantages of those two types of imaging sensors, infrared and visible image fusion (IVIF) technology is proposed, which can combine the mutual information extracted by the two types of sensors into one image with both clear target and background, achieving a more comprehensive description of the scene. Currently, this technology has been widely used in space exploration, resource exploration, biomedicine, equipment detection, target tracking and other fields.

In recent years, a series of research achievements have been made in infrared and visible image fusion. Effective fusion methods have emerged in large numbers, which can be roughly divided into multi-scale transform-based [2]–[5], sparse representation-based [6]–[9], and deep learning-based fusion methods [10]–[13]. Fusion methods based on multi-scale transform generally perform multi-scale transformation on the source images, then fuse the transform coefficients, and use multi-scale inverse transformation to reconstruct the fusion result. However, multi-scale transformation involved in this type of method often represents the image with a fixed basis, which has weak sparsity and limits the further improvement of fusion performance. In contrast, the sparse representation-based fusion methods can effectively alleviate the above problems by building an over complete dictionary for the represented images from a set of training samples. Nevertheless, it still has weak performance in mining the statistical characteristics of large samples. Deep learning-based methods enjoy high favor among researchers lately because they can effectively solve this problem.

At present, numerous excellent deep learning-based IVIF methods have emerged, which are all committed to transferring complementary information from infrared and visible images to the fusion results. Achieving this goal generally requires to make two efforts. On the one hand, the feature extraction network is expected to have excellent performance in complementary information extraction. However, it is extremely challenging for the feature extraction network to effectively perceive complementary information and pay more attention to it. On the other hand, complex fusion strategies are often designed to aggregate the complementary information. In addition, existing fusion schemes all regard harmonious information of the source images as meaningless, and strive to ignore it. Different from the above methods, in order to alleviate challenge posed by the complementary information excavation of feature extraction network, this paper rethinks the deep learning-based fusion idea, that is, complementary to harmonious information transfer, which can effectively avoid weak feature loss of the source images, and transfer those

Y. Zhang, K. Du, H. Li and Z. Yu are with the Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China.(E-mail:lhfchina99@kust.edu.cn (H. Li))

Y. Liu is with Department of Biomedical Engineering, Hefei University of Technology, Hefei, 230009, China.
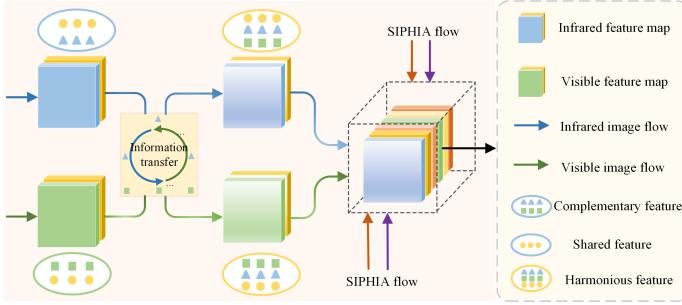
Fig. 1. The core idea of our proposed CHITNet. Harmonious feature indicates the shared information of infrared and visible images and modality-specific information from both two modal source images.

weak details to the fusion results. Specifically, the core idea of the proposed method is shown in Fig. 1. We consider that the infrared and visible images of the same scene cover both shared and modality-specific (complementary) features. If the infrared-specific features are transferred into the visible one, we can obtain features containing shared, infrared-specific and visible-specific information, *i.e.* harmonious features, and vice versa. Compared to the complementary feature based fusion, harmonious feature based fusion can not only reduce the burden of feature extraction network mining complementary information, but also avoid the complementary information loss in the fusion process, getting rid of the reliance on complex fusion schemes.

Based on the above ideas, we propose a dedicated complementary to harmonious information transfer network (CHIT-Net) consisting of mutual information transfer (MIT) module, structure information preserved harmonious information acquisition (SIPHIA) module and fusion result reconstruction (FRR) module. MIT is composed of three parts: image inversion (II), feature extraction (FE) and mutual representation learning (MRL). II mainly facilitates the feature extraction network to obtain richer information. FE is responsible for extracting features from the input images for subsequent information transfer and fusion. MRL realizes interactive information transmission between infrared and visible source images, achieving complementary to harmonious information transfer. SIPHIA contains harmonious information acquisition supervised by source image (HIASSI) module and structure information preservation (SIP) module. HIASSI ensures that the complementary information after MRL can become harmonious one. SIP promises that the edge structure information of the source images can be transferred to the fusion result. FRR constructs the final fusion image. The above design can effectively alleviate complementary feature extraction challenge, and meanwhile avoid designing complex fusion mechanism to aggregate complementary information. Furthermore, similar to [14], we adopt a mutual promotion training paradigm (MPTP) between two SIPHIAs and the MIT. Concretely, we divide the whole training process to two circulative phases, in other words they are trained in turn. In the first cycle phase, we train MIT to have the basic ability of information transfer by performing interactive feature representation. And then in the second cycle phase, we train two SIPHIAs to create impressive edge information for fusion, treating the fusion results obtained by MIT as "labels" to further improve the harmonious feature and edge details extraction ability of two SIPHIAs. In summary, the main contributions of the proposed method are as follows:

- We reconsider the IVIF task from a new aspect, and propose an innovative idea, which transfers complementary information into harmonious one to avoid the tricky complementary information extraction and aggregation problem.
- Benefited from this interesting notion, we exploit an efficient IVIF model CHITNet to realize complementary to harmonious information transfer in source infrared and visible images. Meanwhile, a feature protection mechanism is implanted, migrating detailed edge information to the fusion result.
- Extensive experiments conducted on three publicly available datasets demonstrate the effectiveness of our method, as well as show its superiority over other state-of-the-art (SOTA) models.

The rest of this article is organized as follows: Section II discusses typical deep learning based and high-level task-driven IVIF methods. In Section III, we describe the details of the proposed method. Section IV shows the experimental results and analysis, and Section V summarizes the work and draws some conclusions.

II. RELATED WORK

The IVIF task has been extensively explored for years and attracted much attention from the community due to its wide application prospects. In this section, we will briefly review the existing deep learning based IVIF algorithms that are closely related to ours and some high-level task-driven IVIF algorithms.

A. Deep learning based infrared-visible image fusion

CNN-based methods. In recent years, convolution neural network (CNN) dominates the field of image fusion in virtue of its strong adaptability. Commonly, the entire image fusion procedure includes feature extraction, feature fusion and image reconstruction. Some methods merely use CNN to realize activity level measurement and generate a weight map for hand-crafted features [15], [16], while the main fusion process remains traditional ways. Other methods adopt end-to-end mode, carrying out every step in image fusion with CNN. For instance, Zhang et al. [17] explored a fast unified image fusion network based on proportional maintenance of gradient and intensity. They defined a unified loss function which can adapt to different fusion tasks according to hyperparameters. In order to better preserve the thermal targets in infrared images and the texture structures in visible images, Ma et al. [18] introduced a salient target mask to annotate desired regions. Their network can achieve salient target detection and fuse key information in an implicit manner. Moreover, Xu et al. [19] integrated different fusion tasks into a unified framework with adaptive information preservation degrees.

AE-based methods. Given that the auto-encoder (AE) structure can learn effective feature representation in an unsupervised manner, AE-based methods are utterly suitable for IVIF task since there commonly lacks fused ground-truth. Most of them trained an auto-encoder to achieve both feature extraction and image reconstruction. Typically, Li et al. [20] presented a novel deep learning architecture with encoding network combined by convolutional layers, fusion layer and dense block, while the fused image was reconstructed by decoder. Furthermore, they introduced NestFuse [21] and RFN-Nest [22] to preserve significant amounts of information from input data in a multi-scale perspective. Specifically, the latter designed the detail preservation and feature enhancement loss functions to integrate more useful information from source images. To address the limitation of the uninterpretability of deep feature maps, Xu et al. [23] proposed an unsupervised deep learning method to realize the interpretable importance evaluation of feature maps.

GAN-based methods. Thanks to the strong ability to estimate probability distributions, the generative adversarial network (GAN) is ideal for unsupervised tasks such as image fusion. Ma et al. [24] played an adversarial game between a generator and a discriminator for the first time, keeping the infrared thermal radiation information and preserving the visible appearance texture information. Unfortunately, a single discriminator is prone to lead the fused image to be similar to only one of the source images, causing the loss of some information. Therefore, DDcGAN [25], a dual-discriminator conditional generative adversarial network was designed to simultaneously keep the most important feature information in infrared and visible images. Besides, the existing GAN-based infrared and visible image fusion methods fail to highlight the typical regions in infrared and visible images because of the inability to perceive the most discriminative parts. To this end, AttentionFGAN [26] integrated multi-scale attention mechanism into both generator and discriminator of GAN to capture more typical regions. Subsequently, GANMcC [27] was devised to transform image fusion into muti-distribution simultaneous estimation problem, which made the fused results have both the distributions of infrared and visible domains in a more balanced manner.

Nevertheless, above mentioned deep learning-based approaches put emphasis on extracting and integrating the complementary information from the infrared and visible images, which is an extremely hard problem. To this end, we rethink the IVIF task from a brand new aspect, focusing more on the harmonious information excavation which is relatively easier to achieve.

B. High-level task-driven infrared-visible image fusion

Low-level vision tasks are often pre-processed and serve for downstream high-level vision tasks, *e.g.* object detection, semantic segmentation, *etc.*. For the infrared and visible image fusion task, Tang et al. [28] introduced the semantic loss to integrate more semantic information into fused images. They further incorporated image registration, image fusion, and semantic requirements of high-level vision tasks into a single framework and proposed a novel image registration and fusion method, named SuperFusion [29]. Consequently, since object detection focuses on sparse object regions in images and is important in disaster rescue and traffic field, Sun et al. [30] used detection-driven losses to provide task-related information for the detection-driven IVIF network. More recently, Liu et al. [31] proposed a bilevel optimization formulation for the joint problem of fusion and detection, named TarDAL network for fusion and a commonly used detection network.

However, high-level vision task-driven image fusion often requires elaborately designed down-stream network to tackle the high-level vision task, which is relatively cumbersome and lacking in generality. Conversely, our proposed method is more universal without specific down-stream task oriented framework design.

III. THE PROPOSED METHOD

A. Overview

Most existing IVIF methods are dedicated to mining the complementary information in two source images, which is quite challenging. Therefore, we rethink the problem from another new perspective, where we focus on mining the harmonious information in infrared and visible images, since such information is relatively easy to acquire and utilize in sequential fusion process. To be more specific, we propose a specialized complementary to harmonious information transfer framework to tactfully evade the complementary information extraction problem during infrared and visible image fusion.

Pipeline of the proposed CHITNet is shown in Fig. 2, which consists of the mutual information tranfer (MIT) module, the structure information preserved harmonious information acquisition (SIPHIA) module and the fusion result reconstruction (FRR) module. MIT mutually represents features from two modalities, roughly transferring complementary information into harmonious one. SIPHIA aims to extract harmonious as well as abundant information, especially edge structure details from source images of two modalities. Besides, it is noteworthy that we do not adopt the end-to-end training paradigm but choose muti-stage mutual promotion training strategy to achieve effective interaction and co-reinforcement between MIT and SIPHIA.

B. Mutual information transfer

We design the mutual information transfer (MIT) module to realize complementary to harmonious information transfer. MIT consists of three key components, namely image inversion (II), feature extraction (FE), and mutual representation learning (MRL). Below we describe the concrete structures of II, FE and MRL.

1) Image inversion and feature extraction: To make better use of the content in relatively over-exposed areas, literature [32] inverted the source image so that the originally over-exposed regions would appear like underexposed ones in the inverted image. We find that the salient targets in grayscale inverted image tend to resemble that in the corresponding image from another modality, *i.e.*, the salient target in inverted
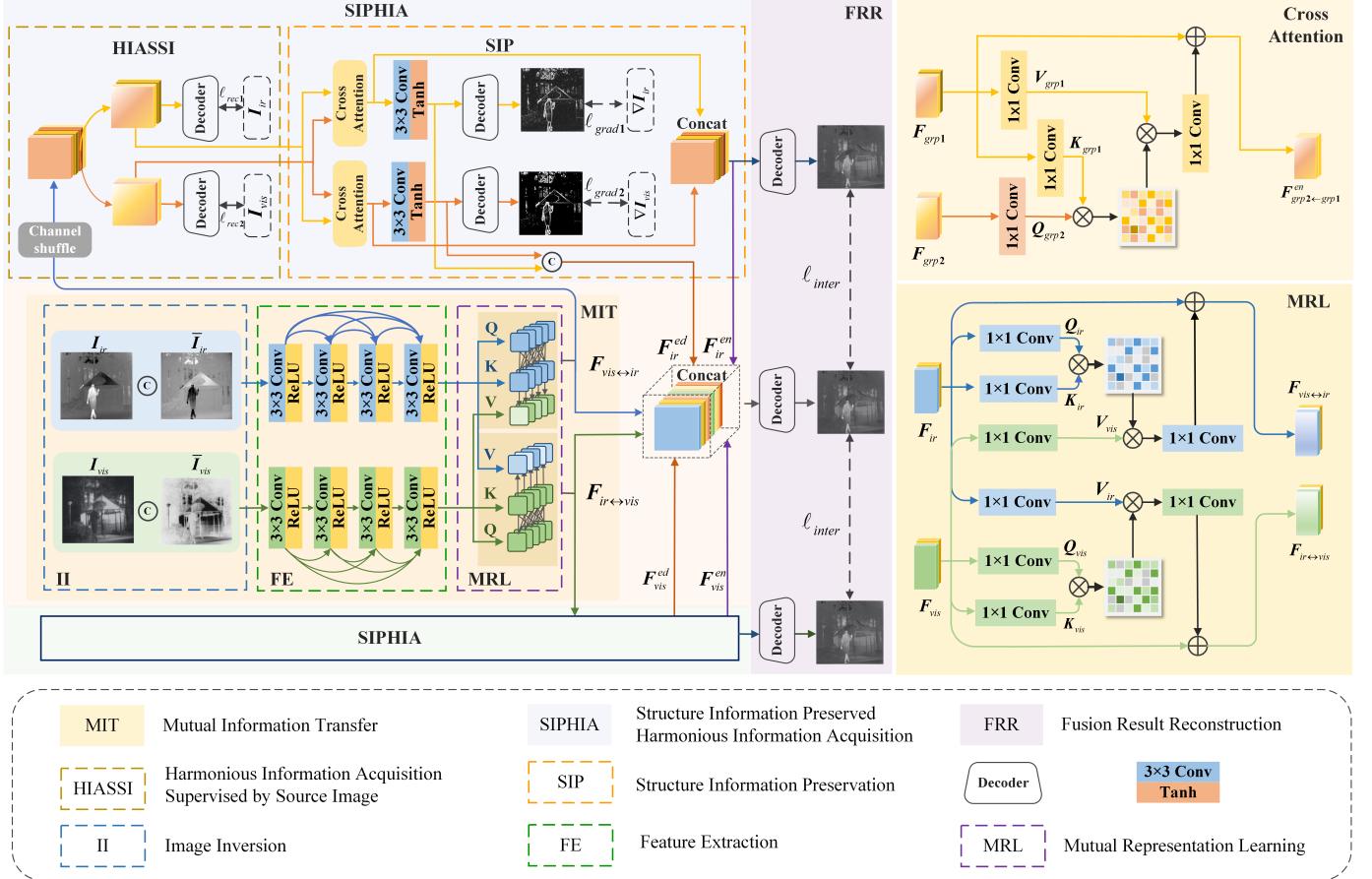
Fig. 2. The overall architecture of proposed method. In phaseM, the input concatenated infrared and visible image pairs I'_{ir} and I'_{vis} are fed into the infrared-visible feature encoder E_{ir} and E_{vis} respectively to obtain features F_{ir} and F_{vis}. MIT is performed on F_{ir} and F_{vis} to achieve basic complementary to harmonious information transfer, getting transferred features $F_{vis\leftrightarrow ir}$ and $F_{ir\leftrightarrow vis}$. In phaseS, we then send $F_{vis\leftrightarrow ir}$ and $F_{ir\leftrightarrow vis}$ to SIPHIA and attain F^{ed}_{ir}, F^{en}_{ir}, F^{ed}_{vis}, and F^{en}_{vis}, ensuring successful information transfer and effectively preserving structure information. Finally, $F_{vis\leftrightarrow ir}$, F^{ed}_{ir}, F^{en}_{ir}, $F_{ir\leftrightarrow vis}$, F^{ed}_{vis}, and F^{en}_{vis} are concatenated together and sent to the decoder D_{fuse} to obtain the final fusion result I_{fused}.

infrared image is alike with that in its corresponding visible image, and so does the inverted visible image. Thus, concatenating the original image and its inverted image will facilitate the complementary to harmonious information transfer. Given source images I_{ir} and I_{vis}, we invert them by $\overline{I}_{ir} = 1 - I_{ir}$ and $\overline{I}_{vis} = 1 - I_{vis}$. Pairs of I_{ir} and $\overline{I}_{ir}$ are concatenated in channel dimensions to obtain I'_{ir}, and pairs of I_{vis} and $\overline{I}_{vis}$ are concatenated to obtain I'_{vis}. I'_{ir} and I'_{vis} are fed into the infrared-visible feature encoder E_{ir} and E_{vis}, respectively. The extracted infrared and visible image features are presented by

$$
\begin{aligned}
F_{ir} &= E_{ir}(I'_{ir}), \\
F_{vis} &= E_{vis}(I'_{vis}),
\end{aligned}
\tag{1}
$$

where E_{ir} and E_{vis} have the same structure but do not share parameters, which follows the residual dense network [33].

2) Mutual representation learning: The architecture of MRL module is presented in Fig. 2, which strives to transfer complementary information hidden in source images to harmonious one through mutual modality feature representation.

Similar to attention mechanism in Transformer [34], the infrared image feature $F_{ir} \in \mathbb{R}^{C \times H \times W}$ (where H and W are the height and width, C is the number of channels) and the

visible image feature $F_{vis} \in \mathbb{R}^{C \times H \times W}$ pass through convolutions to generate Query $Q_{ir} \in \mathbb{R}^{C \times H \times W}$, Key $K_{ir} \in \mathbb{R}^{C \times H \times W}$, Value $V_{ir} \in \mathbb{R}^{C \times H \times W}$, and Query $Q_{vis} \in \mathbb{R}^{C \times H \times W}$, Key $K_{vis} \in \mathbb{R}^{C \times H \times W}$, Value $V_{vis} \in \mathbb{R}^{C \times H \times W}$, respectively. Given F_{ir} and F_{vis}, their mutual feature representations are obtained through information transfer:

$$
\begin{aligned}
\widetilde{F}_{vis\leftrightarrow ir} &= \text{softmax}\left(\frac{Q_{ir}(K_{ir})^T}{\sqrt{d}}\right)V_{vis}, \\
\widetilde{F}_{ir\leftrightarrow vis} &= \text{softmax}\left(\frac{Q_{vis}(K_{vis})^T}{\sqrt{d}}\right)V_{ir},
\end{aligned}
\tag{2}
$$

where $\leftrightarrow$ is cross-modality mutual representation, $\sqrt{d}$ is a normalization factor, and T is transpose operation. To further enhance the gained interactive feature representation and avoid information loss, we perform the following operations:

$$
\begin{aligned}
F_{vis\leftrightarrow ir} &= Conv_{1\times 1}(\widetilde{F}_{vis\leftrightarrow ir}) \oplus F_{ir}, \\
F_{ir\leftrightarrow vis} &= Conv_{1\times 1}(\widetilde{F}_{ir\leftrightarrow vis}) \oplus F_{vis},
\end{aligned}
\tag{3}
$$

where $Conv_{1\times 1}$ is 1×1 convolution. $\oplus$ denotes summing operation.

C. Structure information preserved harmonious information acquisition

Infrared SIPHIA generates F_{ir}^{en} and F_{ir}^{ed}. Similarly, visible SIPHIA generates F_{vis}^{en} and F_{vis}^{ed}. F_m^{en} ($m \in ir, vis$) are enhanced features added to the final fusion process, avoiding information loss, while F_m^{ed} are edge detail features injected to the fused results to help with the structure information preservation. HIPRIA mainly consists of HIASSI and SIP, which will be demonstrated in detail as follows.

1) Harmonious information acquisition supervised by source image: To further ensure the complementary to harmonious information transfer after mutual representation learning, HIASSI is proposed to enable the features injected to the fusion process to have abundant information from both infrared and visible modalities by elaborately designing image reconstruction, so as to maximize the redundancy of the whole network. The specific structure of HIASSI is shown in Fig. 2. $F_{vis \leftrightarrow ir}$ and $F_{ir \leftrightarrow vis}$ are fed into infrared and visible HIASSI, respectively. Take infrared HIASSI as example, to guarantee that the complementary information is transferred to harmonious one after MRL, we first shuffle all the channels of $F_{vis \leftrightarrow ir}$, and then divide the first half of $F_{vis \leftrightarrow ir}$ by channel as F_{grp1}, the rest half as F_{grp2}. The images $\hat{I}_{grp2}$ and $\hat{I}_{grp2}$ are respectively reconstructed with F_{grp1} and F_{grp2} through decoders D_1^{rec} and D_2^{rec} (composed of 3×3 convolution and *Tanh* activation function), whose network structures are the same but with different parameters. To encourage F_{grp1} and F_{grp2} to cover more modality-harmonious information, the reconstruction loss is defined as:

$$\begin{aligned}
\ell_{rec1} &= ||\hat{I}_{grp1} - I_{ir}||_1, \\
\ell_{rec2} &= ||\hat{I}_{grp2} - I_{vis}||_1, \\
\ell_{rec} &= \ell_{rec1} + \ell_{rec2}.
\end{aligned} \quad (4)$$

2) Structure information preservation: As aforementioned, edge detail information plays an important role in the infrared and visible image fusion since it can facilitate subsequent processing and help in decision making. Therefore, we design SIP module to preserve rich edge features, whose structure is displayed in Fig. 2. Firstly, cross attention is performed on both F_{grp1} and F_{grp2} together to obtain interactive and enhanced features. Mathematically, the interactive enhancement process is denoted as:

$$\begin{aligned}
\widetilde{F}_{grp2 \leftarrow grp1} &= \text{softmax}(\frac{Q_{grp2}(K_{grp1})^T}{\sqrt{d}})V_{grp1}, \\
\widetilde{F}_{grp1 \leftarrow grp2} &= \text{softmax}(\frac{Q_{grp1}(K_{grp2})^T}{\sqrt{d}})V_{grp2},
\end{aligned} \quad (5)$$

where $\leftarrow$ denotes feature interaction, $Q_{grp1}, Q_{grp2} \in \mathbb{R}^{C \times H \times W}$, $K_{grp1}, K_{grp2} \in \mathbb{R}^{C \times H \times W}$, and $V_{grp1}, V_{grp2} \in \mathbb{R}^{C \times H \times W}$ are the results of F_{grp1} and F_{grp2} passed through 1×1 convolution, respectively. $\sqrt{d}$ is a normalization factor, and T is transpose operation. To further enhance the features, the following operations are performed:

$$\begin{aligned}
F_{grp2 \leftarrow grp1}^{en} &= Conv_{1 \times 1}(\widetilde{F}_{grp2 \leftarrow grp1}) \oplus F_{grp1}, \\
F_{grp1 \leftarrow grp2}^{en} &= Conv_{1 \times 1}(\widetilde{F}_{grp1 \leftarrow grp2}) \oplus F_{grp2}.
\end{aligned} \quad (6)$$

Secondly, we apply two encoders E_1^{ed} and E_2^{ed} composed of 3×3 convolution and *Tanh* activation function to $F_{grp2 \leftarrow grp1}^{en}$ and $F_{grp1 \leftarrow grp2}^{en}$ so as to obtain edge features $F_{grp2 \leftarrow grp1}^{ed}$ and $F_{grp1 \leftarrow grp2}^{ed}$, which are expected to capture more edge-related information:

$$\begin{aligned}
F_{grp2 \leftarrow grp1}^{ed} &= E_1^{ed}(F_{grp2 \leftarrow grp1}^{en}), \\
F_{grp1 \leftarrow grp2}^{ed} &= E_2^{ed}(F_{grp1 \leftarrow grp2}^{en}).
\end{aligned} \quad (7)$$

To ensure that the network can extract rich edge details, $F_{grp2 \leftarrow grp1}^{ed}$ and $F_{grp1 \leftarrow grp2}^{ed}$ are fed into decoders D_1^{ed} and D_2^{ed} (composed of 3×3 convolution and *Tanh* activation function) to attain the reconstructed images $\hat{I}_1^{ed}$ and $\hat{I}_2^{ed}$. The gradient loss is performed to force E_1^{ed} and E_2^{ed} to extract rich edge features:

$$\begin{aligned}
\ell_{grad1} &= ||\nabla \hat{I}_1^{ed} - \nabla I_{ir}||_1, \\
\ell_{grad2} &= ||\nabla \hat{I}_2^{ed} - \nabla I_{vis}||_1, \\
\ell_{grad} &= \ell_{grad1} + \ell_{grad2},
\end{aligned} \quad (8)$$

where ∇ denotes Laplacian gradient operator.

At this point, we can obtain the edge features extracted by SIP, which can be expressed as:

$$F^{ed} = \left[F_{grp1 \leftarrow grp2}^{ed}, F_{grp2 \leftarrow grp1}^{ed} \right], \quad (9)$$

where $[\cdot]$ denotes channel-wise concatenation. For infrared and visible SIPHIA, F_{ir}^{ed} and F_{vis}^{ed} denote the edge features from them, respectively.

Furthermore, since information loss is inevitable in feature extraction, we also add those enhanced features with rich information to the main fusion procedure to ensure more informative fused results. The enhanced feature in SIPHIA can be expressed as:

$$F^{en} = \left[F_{grp1 \leftarrow grp2}^{en}, F_{grp2 \leftarrow grp1}^{en} \right]. \quad (10)$$

F_{ir}^{en} and F_{vis}^{en} denote the enhanced features from infrared and visible SIPHIA, respectively.

D. Fusion result reconstruction

To transfer complementary information to harmonious information, we present the above MIT module to achieve mutual feature representation betweeen two complementary modalities images, and two SIPHIAs to ease MIT of transferring information from different modalities, together for more visually pleasant fusion results. As for the fusion process, we concatenate the mutual represented features of different modalities $F_{vis \leftrightarrow ir}$ and $F_{ir \leftrightarrow vis}$ in MIT and those acquired useful harmonious information from infrared and visible SIPHIA, i.e., F_{ir}^{ed}, F_{ir}^{en}, F_{vis}^{ed}, and F_{vis}^{ed}, formulated as follows:

$$I_{fused} = D_{fuse}(\left[F_{vis \leftrightarrow ir}, F_{ir}^{ed}, F_{ir}^{en}, F_{ir \leftrightarrow vis}, F_{vis}^{ed}, F_{vis}^{en} \right]), \quad (11)$$

where D_{fuse} stands for the decoder made up of 3×3 convolution and *Tanh* activation.

To retain clear edge details and rich texture information in the fused image, we adopt joint gradient loss ℓ_{JGrad}. The gradient of the fused image is forced to approach the maximum

value between the infrared and visible image gradients. ℓ_{JGrad} is formulated as:

$$\ell_{JGrad} = ||\text{O}(\max(|\nabla \boldsymbol{I}_{ir}|, |\nabla \boldsymbol{I}_{vis}|)) - \nabla \boldsymbol{I}_{fused}||_1, \qquad (12)$$

where ∇ is Laplacian gradient operator. $\max(\cdot)$ denotes taking the maximum value. $\text{O}(|x|) = x$ denotes finding the original gradient value before taking its absolute value.

In addition, to retain the saliency targets from the two images, we introduce the intensity loss, which can be expressed as:

$$\begin{aligned} \boldsymbol{\omega}_{ir} &= \boldsymbol{S}_{\boldsymbol{I}_{ir}}/(\boldsymbol{S}_{\boldsymbol{I}_{ir}} - \boldsymbol{S}_{\boldsymbol{I}_{vis}}), \boldsymbol{\omega}_{vis} = 1 - \boldsymbol{\omega}_{ir}, \\ \ell_{int} &= ||(\boldsymbol{\omega}_{ir} \odot \boldsymbol{I}_{ir} + \boldsymbol{\omega}_{vis} \odot \boldsymbol{I}_{vis}) - \boldsymbol{I}_{fused}||_1, \end{aligned} \qquad (13)$$

where $\boldsymbol{S}_{\boldsymbol{I}_{ir}}$ and $\boldsymbol{S}_{\boldsymbol{I}_{vis}}$ denote saliency matrices of $\boldsymbol{I}_{ir}$ and $\boldsymbol{I}_{vis}$, which can be computed according to references [31], [35]–[37]. $\boldsymbol{\omega}_{ir}$ and $\boldsymbol{\omega}_{vis}$ denote the weight maps for $\boldsymbol{I}_{ir}$ and $\boldsymbol{I}_{vis}$, respectively. $\odot$ denotes element-wise multiplying operation.

The overall reconstruction loss in MIT is computed by

$$\ell_{mit} = \ell_{int} + \lambda_{JG}\ell_{JGrad}, \qquad (14)$$

where λ_{JG} is the hyper-parameter.

For infrared SIPHIA, $\boldsymbol{F}_{ir}^{en}$ is fed into the decoder $\boldsymbol{D}_{ir}^{en}$ (composed of 3×3 convolution and *Tanh* activation function) to obtain reconstructed image $\hat{\boldsymbol{I}}_{ir}^{siphia}$. We use the content loss similar with Eq. (13) to ensure abundant valuable information in enhanced features:

$$\ell_{content}^{en} = ||(\boldsymbol{\omega}_{ir} \odot \boldsymbol{I}_{ir} + \boldsymbol{\omega}_{vis} \odot \boldsymbol{I}_{vis}) - \hat{\boldsymbol{I}}_{ir}^{siphia}||_1. \qquad (15)$$

To encourage the restoration of texture details, we model the gradient distribution and develop an edge loss similar with Eq. (12) as

$$\ell_{edge}^{en} = ||\text{O}(\max(|\nabla \boldsymbol{I}_{ir}|, |\nabla \boldsymbol{I}_{vis}|)) - \nabla \hat{\boldsymbol{I}}_{ir}^{siphia}||_1. \qquad (16)$$

The overall enhancement loss is

$$\ell_{en} = \ell_{content}^{en} + \lambda_{edge}\ell_{edge}^{en}, \qquad (17)$$

where λ_{edge} is the hyper-parameter to balance the content and edge losses. To sum up, the total loss in each SIPHIA can be computed by

$$\ell_{siphia} = \ell_{rec} + \ell_{grad} + \ell_{en}. \qquad (18)$$

E. Mutual promotion training paradigm

Since the mutual represented features from MIT, and edge details, enhanced features from two SIPHIAs all have great influence on the final fusion results, making them collaborate with each other to achieve mutual reinforcement becomes a vital step. Simply, we divide the whole training process to two circulative phases, namely phaseM and phaseS.

In the first phaseM, we train MIT to have the basic complementary to harmonious information transfer and fusion ability. In the first phaseS, two SIPHIAs are trained to create impressive supplementary information, *i.e.* $\boldsymbol{F}_m^{ed}$ and $\boldsymbol{F}_m^{en}$ ($m \in \{ir, vis\}$), which is also the important guarantee of MIT. It is noteworthy that in the next phaseM, MIT realizes fusion based on the transferred features and injected features from two SIPHIAs in the previous phaseS. In this phaseM, we fix the parameters in both two SIPHIAs which have

been optimized before and train MIT. As for phaseS training, once the network in MIT get trained, we fix the optimized parameters of MIT trained in earlier phaseM, and take the fused results $\boldsymbol{I}_{fused}$ as "labels", pulling close "labels" and two enhanced harmonious fusion results $\hat{\boldsymbol{I}}_{ir}^{siphia}$ and $\hat{\boldsymbol{I}}_{vis}^{siphia}$ through Kullback-Leibler Divergence, formulated as:

$$\ell_{inter} = \text{KL}(\boldsymbol{I}_{fused}||\hat{\boldsymbol{I}}_{ir}^{siphia}) + \text{KL}(\boldsymbol{I}_{fused}||\hat{\boldsymbol{I}}_{vis}^{siphia}). \qquad (19)$$

In this way, the KL Divergence forces networks in two SIPHIAs to learn better edge details and enhanced features injected to the fusion process, getting added information with higher quality. In the next training stage of MIT, the fused results will be naturally improved with the help of better additional information consequently. By now, a virtuous mutual promotion training cycle has been formed. We further demonstrate the pseudo code of our proposed training paradigm as follows.

Algorithm 1 Mutual Promotion Training Paradigm

Input: infrared images $\boldsymbol{I}'_{ir}$ and infrared images $\boldsymbol{I}'_{vis}$.
Output the fused image $\boldsymbol{I}_{fused}$

1: Sample a batch of source data.
2: Initialize the related encoders, **Q**, **K**, **V** and decoders in MIT and SIPHIAs.
3: **for** $iter$=1, $\cdots$, $Iteration_{Maximum}$ **do**
4:   **if** iter% 200 = even number (phaseM)
5:     Train MIT.
6:     Update $\boldsymbol{E}_{ir}$, $\boldsymbol{E}_{vis}$, $\mathbf{Q_{ir}}$, $\mathbf{K_{ir}}$, $\mathbf{V_{ir}}$, $\mathbf{Q_{vis}}$, $\mathbf{K_{vis}}$, $\mathbf{V_{vis}}$ and $\boldsymbol{D}_{fuse}$ by minimizing the loss in Eq. (14).
7:   **else** (phaseS)
8:     Load the learned $\boldsymbol{E}_{ir}$, $\boldsymbol{E}_{vis}$, $\mathbf{Q_{ir}}$, $\mathbf{K_{ir}}$, $\mathbf{V_{ir}}$, $\mathbf{Q_{vis}}$, $\mathbf{K_{vis}}$, $\mathbf{V_{vis}}$.
9:     Train two SIPHIAs.
10:     Update $\boldsymbol{D}_1^{rec}$, $\boldsymbol{D}_2^{rec}$, $\mathbf{Q_{grp1}}$, $\mathbf{K_{grp1}}$, $\mathbf{V_{grp1}}$, $\mathbf{Q_{grp2}}$, $\mathbf{K_{grp2}}$, $\mathbf{V_{grp2}}$, $\boldsymbol{E}_1^{ed}$, $\boldsymbol{E}_2^{ed}$, $\boldsymbol{D}_1^{ed}$, $\boldsymbol{D}_2^{ed}$ and $\boldsymbol{D}_{ir}^{en}(\boldsymbol{D}_{vis}^{en})$ by minimizing the loss in Eq. (18) and Eq. (19).
10:   **end if**
11:**end for**

IV. EXPERIMENTS

A. Experimental configurations

Training dataset: We select 148 image pairs from the RoadScene [19] benchmark as the training set in this paper, containing rich scenes, such as roads, vehicles, and pedestrians. Since such data scale is insufficient to train a network with superior performance, to expand this dataset, we randomly crop the training samples into 120×120 patches, in which way the whole training set is enlarged to 148000 sample pairs.

Testing dataset: Firstly, we conduct qualitative and quantitative experiments on chosen 38 pairs of testing set from the RoadScene. Moreover, as is known to all that the generalization ability of the deep learning-based method is also a vital indicator to evaluate model performance, we carry out the generalization experiments on 40 image pairs of the TNO [38] and the VOT2020-RGBT [39] datasets and 40 image pairs of the OTCBVS [40] dataset including scenes of busy pathway

intersections on the Ohio State University campus to verify the effectiveness and generalization ability of our CHITNet. TNO dataset contains multi-spectral nighttime imagery of various military-relevant scenarios in grayscale. VOT2020-RGBT is a RGB-infrared paired dataset involving loads of video sequences for Visual-Object-Tracking Challenge.

Evaluation metrics: For quantitative evaluation, six statistical metrics are selected to objectively assess the fusion performance, including correlation coefficient (CC) [41], entropy (EN) [42], $Q^{AB/F}$ [43], Q^{CV} [44], the sum of correlations of differences (SCD) [45], and structural similarity (SSIM) [46]. CC evaluates the degree of linear correlation between the fused image and source images. EN measures the amount of information involved in the fused image on the basis of information theory. $Q^{AB/F}$ calculates how much edge information is transferred from source images to the fused image. Q^{CV} is based on the human visual system model, using the Sobel operator to extract the edge information of the source images and the fusion result to obtain the edge intensity map G. Smaller Q^{CV} implies more in line with human visual perception. SCD reflects the correlation level between information transmitted to the fused image and corresponding source images. SSIM models image loss and distortion. In addition, a fusion algorithm with larger CC, EN, $Q^{AB/F}$, SCD, and SSIM indicates better fusion performance.

B. Implementation details

We train our proposed CHITNet on selected 148 image pairs with data argumentation from the RoadScene dataset. Since our method is a circulative two-stage model, we train each phase with batch size set to 8 and patch size set to 120×120. $Iteration_{Maximum}$ is set to 1000. Furthermore, four Adam optimizers [47]($\beta_1 = 0.9$, and $\beta_2 = 0.999$) responsible for two SIPHIAs and two different MIT stages with initial learning rate of 0.001, which decreases to 10^{-4} after 100 epochs and then to 10^{-7} after 400 epochs, except for the second MIT phase with initial learning rate of 10^{-5} and the same decay rate, are used to optimize two cyclical training phases under the guidance of the interaction loss ℓ_{inter}. In addition, two hyperparameters λ_{edge} and λ_{JG} are set to 20. Our framework is implemented in PyTorch. All experiments are conducted on an Nvidia 3090 GPU.

Please note that we simply convert source images in Road-Scene, VOT2020-RGBT and OTCBVS datasets to gray to achieve fusion.

C. Ablation studies

In the above description, we mainly devise two important modules named MIT and SIPHIA to achieve complementary to harmonious information transfer and edge detail preservation. In this section, we give more results under different settings on the framework and training strategy on TNO-VOT dataset, to verify the rationality of our design.

1) Mutual information transfer analysis: To skillfully achieve IVIF fusion without direct complementary information excavation, we devise the complementary to harmonious

TABLE I
QUANTITATIVE EVALUATION RESULTS OF ABLATION STUDY ON 40 PAIRS OF INFRARED AND VISIBLE IMAGES FROM TNO-VOT DATASET.

Configuration	CC	EN	$Q^{AB/F}$	Q^{CV}	SCD	SSIM
w/o MIT	*0.6874*	6.7817	0.4520	598.0829	1.2261	1.3919
w/o SIPHIA	0.6843	6.9485	*0.4570*	524.1034	*1.3557*	1.3945
w/o MPTP	0.6813	*7.0756*	0.4143	*454.8244*	1.3168	*1.4202*
Ours	**0.6991**	**7.2819**	**0.5503**	**427.2510**	**1.4893**	**1.4443**

(a) IR   (b) VIS   (c) W/O MIT
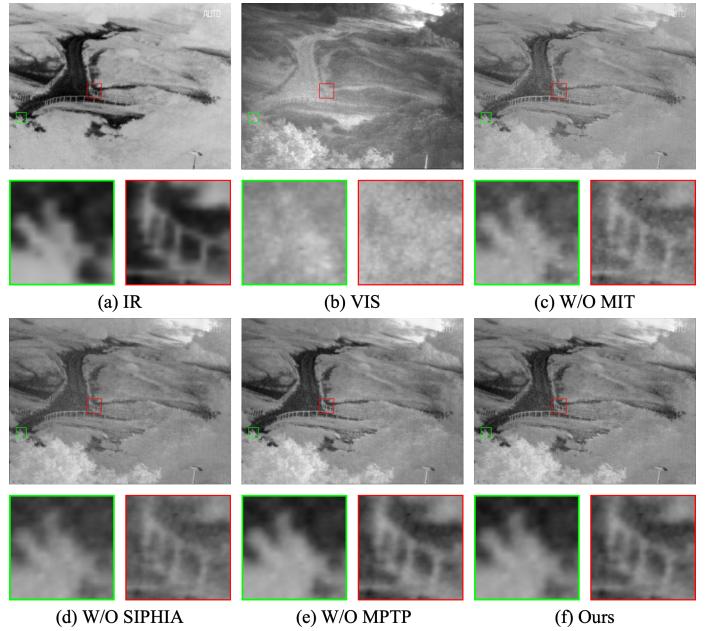
(d) W/O SIPHIA   (e) W/O MPTP   (f) Ours

Fig. 3. Vision quality comparison of the ablation study on important idea and modules. From left to right, infrared image, visible image, and the results of W/O MIT, SIPHIA, MPTP, and our CHITNet.

information transfer to concentrate more on harmonious information. Specifically, we design the MIT module to achieve information transfer, as well as two SIPHIAs to help MIT ensure successful transfer process. After we change the whole network into a single feature extraction and fusion one, *i.e.*, features extracted are modality-specific, the quality of fused image declines slightly. As shown in Fig. 3(c), some edge details of the bush and handrail in the infrared image are lost, and the image contrast decreases. Furthermore, so as to demonstrate the effectiveness of the MIT more intuitively, we display the averaged infrared and visible feature maps on channel before and after the MIT module, *i.e.*, $\boldsymbol{F}_{ir}$, $\boldsymbol{F}_{vis \leftrightarrow ir}$, $\boldsymbol{F}_{vis}$ and $\boldsymbol{F}_{ir \leftrightarrow vis}$ in Fig. 4. We can clearly see that the strip on the road sign in visible image successfully transfers to the infrared image. Moreover, the wire netting and walking man in the infrared image successfully transfer to the visible image, which implies successful complementary to harmonious information transfer. Besides, the quantitative results in Table I drop the most in practically all evaluation metrics, which speaks volumes for our complementary to harmonious information transfer idea.

2) Structure information preserved harmonious information acquisition analysis: Two SIPHIAs are responsible for the
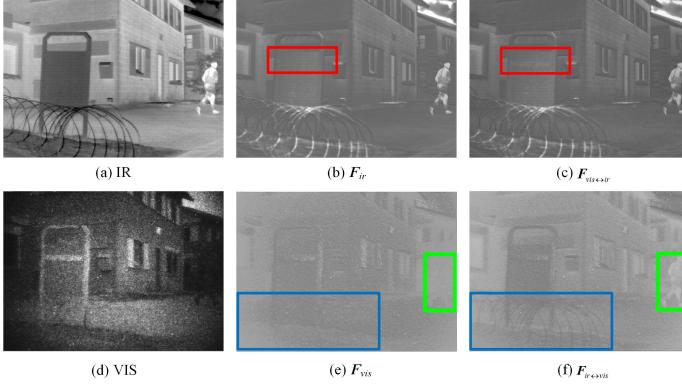
Fig. 4. Infrared and visible feature maps before and after mutual information transfer.

TABLE II
QUANTITATIVE RESULTS OF SEVEN SOTA METHODS AND OURS ON 38 IMAGE PAIRS FROM ROADSCENE [19] DATASET. BOLD: BEST. ITALIC: SECOND BEST.

Methods	CC	EN	$Q^{AB/F}$	Q^{CV}	SCD	SSIM
RFN-Nest [22]	0.5926	*7.3512*	0.2983	1275.9437	1.2189	1.2337
SuperFusion [29]	*0.6719*	7.0620	0.4604	**491.2926**	1.3930	1.3072
FusionGAN [24]	0.5540	7.1681	0.2588	1442.2847	0.8642	1.1566
U2Fusion [19]	0.5847	6.8581	*0.4658*	1071.4871	1.1965	*1.3822*
SDDGAN [48]	0.5601	**7.4797**	0.3273	1319.1305	1.2830	1.1960
MetaFusion [49]	0.6533	7.0634	0.3441	1123.0795	*1.4954*	1.0712
LRRNet [50]	0.5558	6.9726	0.1890	1298.3829	0.5990	0.8598
Ours	**0.6786**	7.1010	**0.4733**	*867.5416*	**1.5388**	**1.3944**

retention and boost of edge and texture, in which HIASSI and SIP modules are also the key points in our overall framework since they can effectively enhance the edge and texture details in the fused images. When we remove both of the two SIPHIAs, the edges of nearly all the objects are indistinguishable, meanwhile the texture details blur, as shown in Fig. 3(d). On the contrary, our fusion result presents more detailed textures as shown in Fig. 3(f). Since $Q^{AB/F}$ measures the amount of edge information that is transferred from source images to the fused image, lower $Q^{AB/F}$ in model without SIPHIAs further indicates the effectiveness of the proposed SIPHIA module.

3) Mutual promotion training paradigm analysis: We propose a circulative mutual promotion training strategy to improve the final fusion performance across the board. In the ablation experiment, we adopt end-to-end training paradigm, which is to say, we train two SIPHIAs and MIT together as a whole. The visualized result is presented in Fig. 3(e). Though the visualized result is similar to ours, the quantitative results in Table I deteriorate. Therefore, we can draw a conclusion that MPTP provides a more effective training for the fusion process.

D. Comparative experiment

To illustrate the superiority of the proposed method, we conduct comprehensive comparative experiments of fusion performance with seven SOTA competitors including RFN-Nest [22], SuperFusion [29], FusionGAN [24], U2Fusion [19], SDDGAN [48], MetaFusion [49], and LRRNet [50] on 38 image pairs of RoadScene, 40 image pairs of the TNO and the VOT2020-RGBT, and 40 image pairs of OTCBVS datasets.

1) Experiments on RoadScene dataset: So as to visually evaluate the fusion performance of different algorithms on RoadScene dataset, two pairs of infrared and visible images (named #FLIR_00288 and #FLIR_08835, respectively) are selected, shown in Fig. 5. As illustrated in the green and red boxes in the #FLIR_00288 image, our proposed CHITNet keeps the clearest cloud edge while maintaining favourable shrub textures. RFN-Nest, FusionGAN and SDDGAN have vague overall scenes, which declines the visual effects. Super-

Fusion and LRRNet lose cloud information, and are unable to present satisfying contrast.

As shown in the green box of the #FLIR_08835 image, fusion results of RFN-Nest, FusionGAN, SDDGAN, and LRRNet are relatively blurred, missing quite a lot of scene information. Although SuperFusion and U2Fusion have more sharpened pedestrain edges, the railing textures are lost, as zoomed in the red boxes. Oppositely, our fusion results obtain both clear target edges and rich texture details.

We conduct quantitative comparisons on 38 image pairs from RoadScene dataset to verify the effectiveness of our method, which is presented in Fig. 6 and Table II. It can be seen that our method ranks first in four metrics and second in the Q^{CV} metric. The best CC, SCD and SSIM metrics demonstrate that our results contain more realistic information. The presence of the SIP module allows the $Q^{AB/F}$ metric of our method to outperform other SOTA methods. Our method trails SuperFusion by a narrow margin in the Q^{CV} metric, but our fusion result preserves more detailed textures and the scene is clearer than that of other methods.

In conclusion, our CHITNet is fully capable of excavating edge and texture features and integrating them into fused images with the help of two SIPHIAs and MIT. Thereby, our method is superior to other SOTA approaches and obtains high-quality fused images.

2) Experiments on TNO-VOT dataset: In order to visually observe the fusion performance of different algorithms on the selected TNO-VOT dataset, we select eight pairs of infrared and visible images. The visualized results are presented in Fig. 7. As shown in the first column of Fig. 7, nearly all methods have blurred wire netting edges while our method keeps clear edge information, better viewed in the zoomed-in red box. Although SuperFusion performs well in retaining texture details of the wires, it loses clouds textures in the sky, best viewed in the green box. The same situation also happens to the fourth and seventh columns of Fig. 7. On the other hand, our method preserves both edge details and background textures to the most extent.

As depicted in the second and third columns of Fig. 7, the highlighted area in the red box indicates that our proposed method could reach the sharpest edges and maintain satisfying brick textures. Moreover, in the sixth column of Fig. 7, fusion results of SuperFusion and LRRNet submerge fence into fuzzy background. Although RFN-Nest, FusionGAN, and SDDGAN

Fig. 5. Vision quality comparison of our method with seven SOTA fusion methods on #FLIR_00288 and #FLIR_08835 images from the RoadScene dataset.

have relatively clearer edges, they are still indistinct. On the contrary, our method reserves sharpened fence edges as well as considerable texture details.

Finally, as presented in the eighth column of Fig. 7, edge details of the saliency target blur in FusionGAN results, largely due to the insufficient capacity in fully modelling the data distributions of source images. Typically, fused results in U2Fusion method lose the thermal target information and cannot attract human attention. Furthermore, existing approaches tend to extract background detail features from the visible image, while acquire salient features from the infrared image. They neglect texture information in the infrared image, resulting partly detail missing. Instead, our proposed method is able to highlight prominent target while preserving favorable contrast and edge details, as well as introducing less noise.

In conclusion, our method could retain the intensity distribution of the infrared target with more sharpened edges and skillfully exclude useless information, *i.e.* noises in visible images. We attribute this advantage to the fact that our method injects useful edge detail features into the fusion process and spontaneously exploits valuable features in source images.

The comparative results are presented in Table III. We can see that our method ranks first in four metrics and second in the CC and Q^{CV} metrics. The best EN, SCD and SSIM metrics indicate that our results are the most similar with the source images, preserving the original information to the

TABLE III
QUANTITATIVE RESULTS OF SEVEN SOTA METHODS AND OURS ON 40 IMAGE PAIRS FROM TNO [38] AND VOT [39] DATASETS. BOLD: BEST. ITALIC: SECOND BEST.

Methods	CC	EN	$Q^{AB/F}$	Q^{CV}	SCD	SSIM
RFN-Nest [22]	0.6989	7.1465	0.3346	652.7588	1.4376	1.3501
SuperFusion [29]	0.6643	6.9682	0.4081	**353.5487**	1.3611	1.4144
FusionGAN [24]	0.6790	6.5426	0.2127	1238.3153	0.7275	1.2593
U2Fusion [19]	**0.7032**	6.6908	*0.4622*	585.6560	1.2861	*1.4423*
SDDGAN [48]	0.6548	7.2675	0.3084	709.7720	1.4151	1.3197
MetaFusion [49]	0.6735	*7.2748*	0.3489	590.1238	*1.4735*	1.0615
LRRNet [50]	0.6749	7.0341	0.3047	976.6413	1.0897	1.2392
Ours	*0.6990*	**7.2819**	**0.5503**	*427.2510*	**1.4893**	**1.4443**

utmost extent and more natural. The idea of complementary to harmonious information transfer promises the advantage of our approach over the SOTA methods. Moreover, our method also performs well in $Q^{AB/F}$, showing that our fusion results have richer texture details and more salient targets, with the help of SIP module. Besides, CHITNet has an excellent performance on the Q^{CV} metric merely following SuperFusion, implying that our fused images are also consistent with the human visual system. More intuitive results comparison is shown in Fig. 8.

3) Experiments on OTCBVS dataset: Qualitative comparison results on OTCBVS dataset are shown in Fig. 9 to testify the generalization ability of our CHITNet. Our method is able
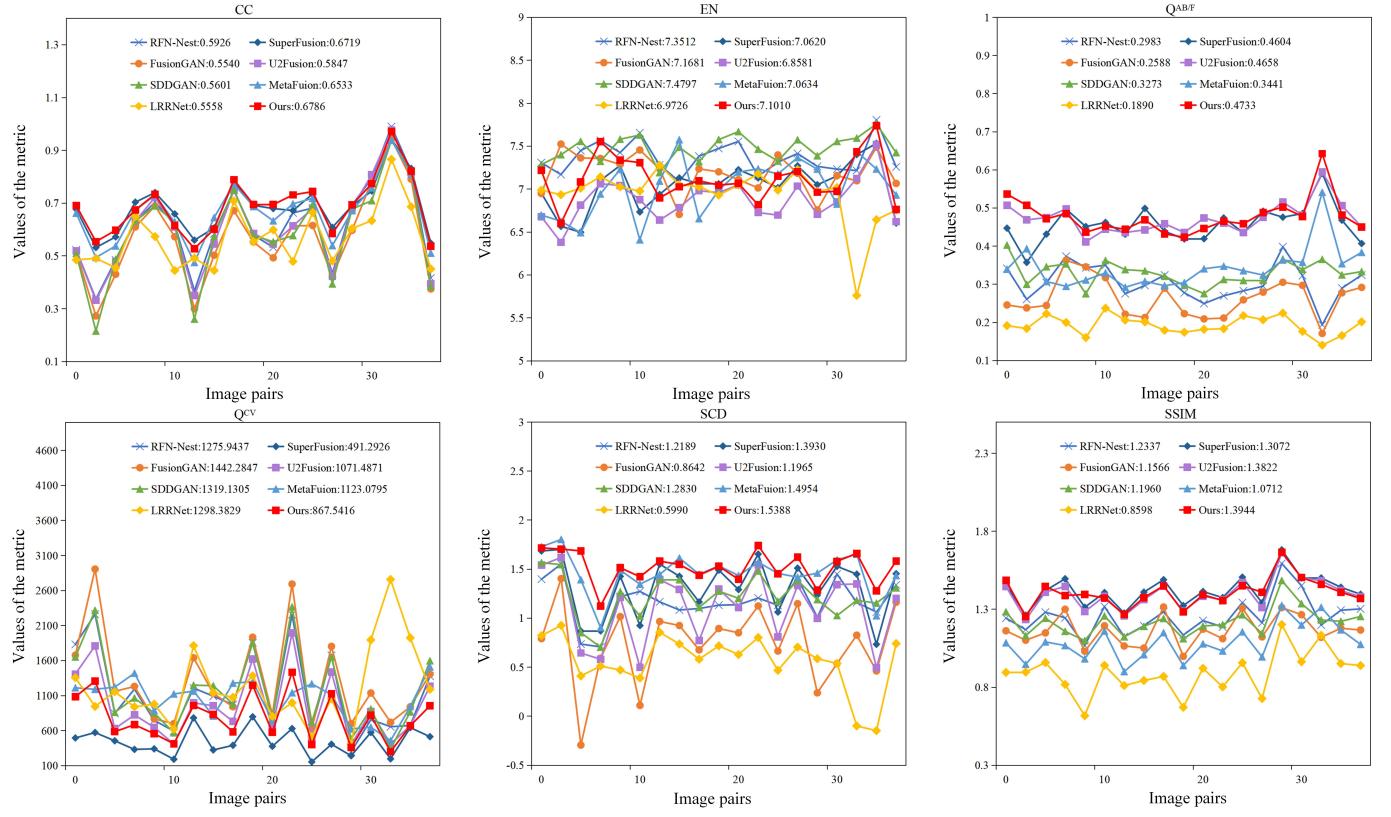
Fig. 6. Quantitative results of six metrics, *i.e.*, CC, EN, $Q^{AB/F}$, Q^{CV}, SCD, and SSIM, on 38 image pairs from RoadScene dataset. Seven SOTA methods are used for comparison.

to provide more sharpened target edges and detailed textures. As shown in the green box of scene #video_0001, edge and texture of the walking man with suit is much clearer in our method while other competitors blur to some extent. What's more, though U2Fusion performs neck and neck with ours, it fails to retain the exact shape of the manhole cover, best viewed in the red zoomed-in boxes.

Table IV gives quantitatively comparable results, and our CHITNet has three indexes ranking first. The best CC, $Q^{AB/F}$, SCD and SSIM metrics imply that our results reserve the most information of the source images and have desirable texture details and salient targets. Besides, the second best in Q^{CV} metric indicates that our fusion results are more natural and human visual-friendly. More intuitively comparable results are shown in Fig. 10.

To sum up, the proposed CHITNet is adept at transferring complementary information to harmonious one, excavating useful information in source images and integrating it into fused images with the help of edge details and enhanced features injection from two SIPHIAs in a co-reinforcement manner. Hence, our method presents more advantages than other SOTA methods and attains high-quality fusion results.

E. Analysis of hyperparameters

We perform hyperparameter analysis on RoadScene dataset [19] to validate the rationality of our CHITNet. As described in

TABLE IV
QUANTITATIVE RESULTS OF SEVEN SOTA METHODS AND OURS ON 40 IMAGE PAIRS FROM OTCBVS [40] DATASET. BOLD: BEST. ITALIC: SECOND BEST.

Methods	CC	EN	$Q^{AB/F}$	Q^{CV}	SCD	SSIM
RFN-Nest [22]	0.5327	7.1992	0.2419	768.3358	1.2989	1.1859
SuperFusion [29]	*0.5450*	*7.4748*	*0.4347*	**672.1874**	1.4110	1.1656
FusionGAN [24]	0.5058	6.3827	0.1678	1401.0553	0.5740	1.1442
U2Fusion [19]	0.5312	7.1555	0.3690	822.4050	1.2984	1.2233
SDDGAN [48]	0.5266	7.4262	0.2887	1216.9237	*1.4413*	*1.2322*
MetaFusion [49]	0.5204	**7.7999**	0.3352	1199.6316	1.4352	0.9548
LRRNet [50]	0.5106	7.1296	0.1667	1492.4342	0.6692	0.8127
Ours	**0.5451**	6.9565	**0.4709**	*766.1982*	**1.4476**	**1.2475**

TABLE V
EFFECT OF DIFFERENT HYPERPARAMETERS SETTINGS ON FUSION PERFORMANCE.

λ_{edge}, λ_{JG}	CC	EN	$Q^{AB/F}$	Q^{CV}	SCD	SSIM
1	0.6335	**7.3042**	0.4802	878.0353	1.2989	1.4926
20	**0.6786**	7.1010	**0.4904**	**867.5416**	**1.5388**	**1.3944**
50	0.6181	7.0921	0.1678	932.8368	0.5740	1.4254
100	0.6006	7.0501	0.4911	987.6895	1.2984	1.3956

SIP module of Section III-C2 and the fusion result reconstruction process in Section III-D, λ_{edge} and λ_{JG} are the parameters controlling the degree of edge loss in the image reconstruction

Fig. 7. Vision quality comparison of our method with seven SOTA fusion methods on eight images from the TNO-VOT dataset. Each column represents a scene.

procedure. According to Eq. (17) and Eq. (14), the larger λ_{edge} and λ_{JG} values pay more attention to the edge maintenance. Since the loss functions for both fused image reconstruction of $\hat{I}_{ir}^{siphia}(\hat{I}_{vis}^{siphia})$ and I_{fused} are the same, we change those two hyperparameters simultaneously. Now that the edge and texture details directly affect the qualities of fusion results, we assign a large value to λ_{edge} and λ_{JG}. In order to analyze the sensitivity of this parameter, λ_{edge} and λ_{JG} are set as 1, 20, 50, and 100, respectively. Vision and numerical quality results are shown in Fig. 11 and Table V. It can be seen that when λ_{edge} and λ_{JG} are set as 20, our fusion result achieves great visual perception and quantitative best.

V. CONCLUSION

In this paper, we rethink the IVIF task from a brand new angle, focusing more on the easily acquired harmonious information rather than struggling and striving on the complementary information extraction. We propose a novel mutual promoted infrared and visible image fusion algorithm with mutual information transfer and structure information preserved harmonious information acquisition, which achieves effective image fusion with more texture retention. Specifically, MIT fulfills the core idea of complementary to harmonious information transfer, HIASSI is designed to ensure that the complementary information is successfully transferred to harmonious one while SIP is for structure information preservation and enhanced feature injection to avoid vital
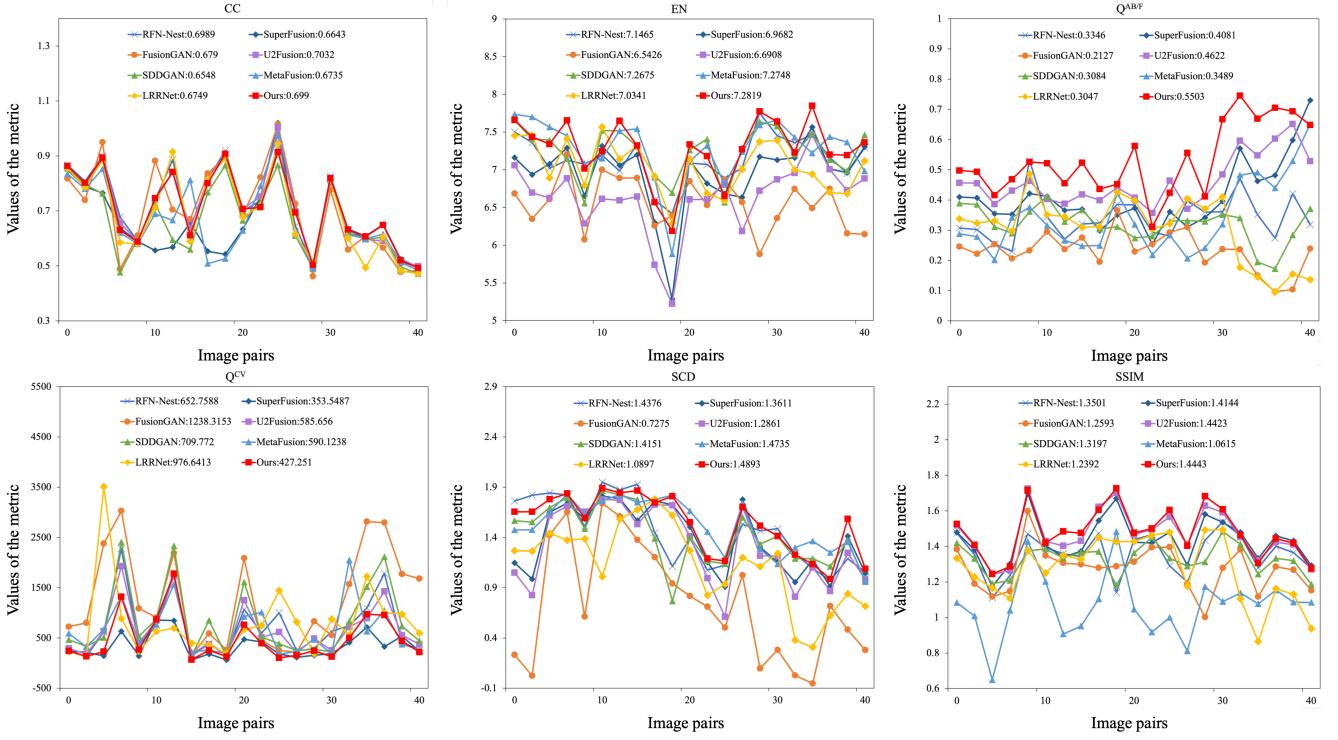
Fig. 8. Quantitative results of six metrics, *i.e.*, CC, EN, $Q^{AB/F}$, Q^{CV}, SCD, and SSIM, on 40 image pairs from TNO-VOT dataset. Seven SOTA methods are used for comparison.

(a) IR (b) VIS (c) RFN-Nest (d) SuperFusion (e) FusionGAN

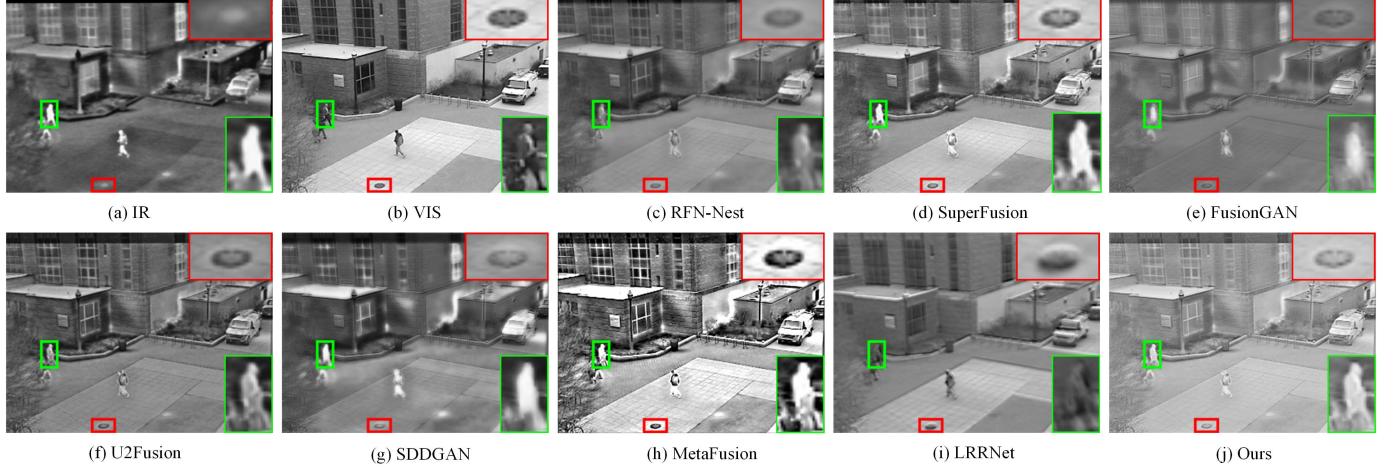(f) U2Fusion (g) SDDGAN (h) MetaFusion (i) LRRNet (j) Ours

Fig. 9. Vision quality comparison of our method with seven SOTA fusion methods on #video_0001 from the OTCBVS dataset.

weak information loss of source images. Then, we employ interaction fusion with the help of added edge details and enhanced features to attain high-quality fused results. Given that components in previous end-to-end training strategies fail to progressively learn from one another, we adopt a mutual promotion training paradigm for better cooperation between each modules. Qualitative and quantitative comparisons with the state-of-the-arts approaches validate the superiority of our method, including edge details, texture information, and visual perception.

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. 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.

[3] H. Li, X. Qi, and W. Xie, "Fast infrared and visible image fusion with structural decomposition," *Knowledge-Based Systems*, vol. 204, p. 106182, 2020.

[4] J. Yao, Y. Zhao, Y. Bu, S. G. Kong, and J. C.-W. Chan, "Laplacian pyramid fusion network with hierarchical guidance for infrared and
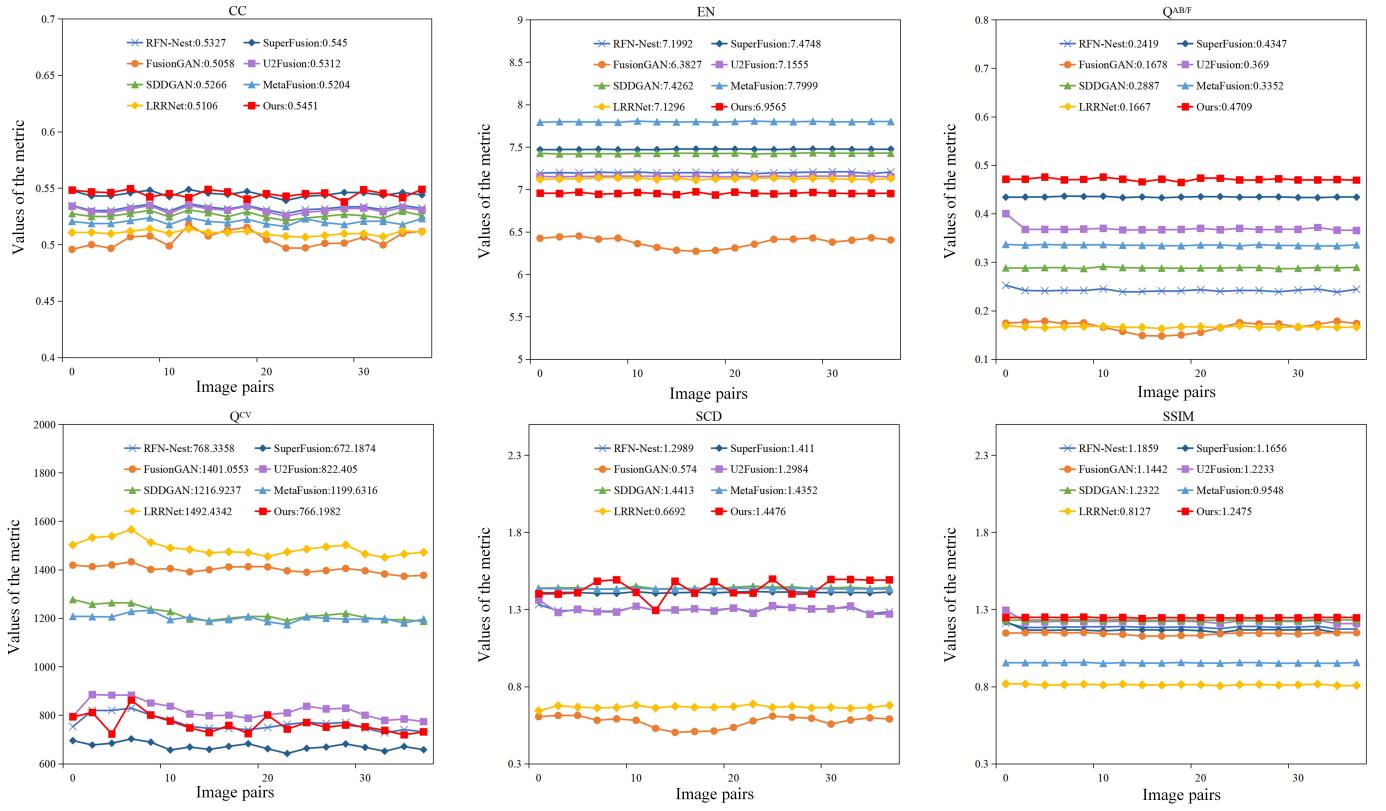
Fig. 10. Quantitative results of six metrics, *i.e.*, CC, EN, $Q^{AB/F}$, Q^{CV}, SCD, and SSIM, on 40 image pairs from OTCBVS dataset. Seven SOTA methods are used for comparison.
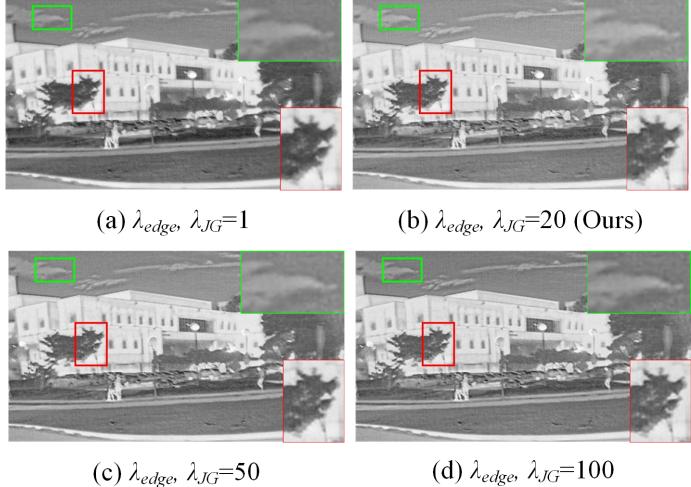

(a) λ_{edge}, λ_{JG}=1

(b) λ_{edge}, λ_{JG}=20 (Ours)

(c) λ_{edge}, λ_{JG}=50

(d) λ_{edge}, λ_{JG}=100

Fig. 11. The fusion results with different λ_{edge} and λ_{JG}.

visible image fusion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 9, pp. 4630–4644, 2023.

[5] X. Yan, H. Qin, J. Li, H. Zhou, and J.-g. Zong, "Infrared and visible image fusion with spectral graph wavelet transform," *JOSA A*, vol. 32, no. 9, pp. 1643–1652, 2015.

[6] L. Sun, Y. Li, M. Zheng, Z. Zhong, and Y. Zhang, "Mcnet: Multiscale visible image and infrared image fusion network," *Signal Processing*, vol. 208, p. 108996, 2023.

[7] M. Xie, J. Wang, and Y. Zhang, "A unified framework for damaged image fusion and completion based on low-rank and sparse decomposition," *Signal Processing: Image Communication*, vol. 29, p. 116400, 2021.

[8] Y. Zhang, M. Yang, NanLi, and Z. Yu, "Analysis-synthesis dictionary pair learning and patch saliency measure for image fusion," *Signal Processing*, vol. 167, p. 107327, 2020.

[9] H. Li, Y. Wang, Z. Yang, R. Wang, X. Li, and D. Tao, "Discriminative dictionary learning-based multiple component decomposition for detail-preserving noisy image fusion," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 4, pp. 1082–1102, 2019.

[10] Z. Zhao, S. Xu, J. Zhang, C. Liang, C. Zhang, and J. Liu, "Efficient and model-based infrared and visible image fusion via algorithm unrolling," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1186–1196, 2022.

[11] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, and Y. Ma, "Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 7, pp. 1200–1217, 2022.

[12] W. Tang, F. He, Y. Liu, Y. Duan, and T. Si, "Datfuse: Infrared and visible image fusion via dual attention transformer," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 7, pp. 3159–3172, 2023.

[13] H. Li, Y. Cen, Y. Liu, X. Chen, and Z. Yu, "Different input resolutions and arbitrary output resolution: A meta learning-based deep framework for infrared and visible image fusion," *IEEE Transactions on Image Processing*, vol. 30, pp. 4070–4083, 2021.

[14] H. Tang, C. Yuan, Z. Li, and J. Tang, "Learning attention-guided pyramidal features for few-shot fine-grained recognition," *Pattern Recognition*, vol. 130, p. 108792, 2022.

[15] H. Li, X. Wu, and J. Kittler, "Infrared and visible image fusion using a deep learning framework," in *24th International Conference on Pattern Recognition*. IEEE, 2018, pp. 2705–2710.

[16] Y. Liu, X. Chen, J. Cheng, and H. Peng, "A medical image fusion method based on convolutional neural networks," in *20th International*

Conference on Information Fusion. IEEE, 2017, pp. 1–7.

[17] H. Zhang, H. Xu, Y. Xiao, X. Guo, and J. Ma, "Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 797–12 804.

[18] 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, pp. 1–13, 2021.

[19] 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, 2020.

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

[21] 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.

[22] 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. 72–86, 2021.

[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, 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.

[25] 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.

[26] 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.

[27] 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, pp. 1–14, 2021.

[28] 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.

[29] L. Tang, Y. Deng, Y. Ma, J. Huang, and J. Ma, "Superfusion: A versatile image registration and fusion network with semantic awareness," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 12, pp. 2121–2137, 2022.

[30] Y. Sun, B. Cao, P. Zhu, and Q. Hu, "Detfusion: A detection-driven infrared and visible image fusion network," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 4003–4011.

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

[32] 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.

[33] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image restoration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 7, pp. 2480–2495, 2020.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[35] S. Ghosh, R. G. Gavaskar, and K. N. Chaudhury, "Saliency guided image detail enhancement," in *2019 National Conference on Communications*. IEEE, 2019, pp. 1–6.

[36] J. Liu, Y. Wu, Z. Huang, R. Liu, and X. Fan, "Smoa: Searching a modality-oriented architecture for infrared and visible image fusion," *IEEE Signal Processing Letters*, vol. 28, pp. 1818–1822, 2021.

[37] D. Wang, J. Liu, X. Fan, and R. Liu, "Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 7 2022, pp. 3508–3515.

[38] A. Toet, "Tno image fusion dataset," 2022. [Online]. Available: https://figshare.com/articles/dataset/TNO_Image_Fusion_Dataset/1008029

[39] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, J.-K. Kämäräinen, M. Danelljan, L. Č. Zajc, A. Lukežič, O. Drbohlav *et al.*, "The eighth visual object tracking vot2020 challenge results," in *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16.* Springer, 2020, pp. 547–601.

[40] J. W. Davis and V. Sharma, "Background-subtraction using contour-based fusion of thermal and visible imagery," *Computer Vision and Image Understanding*, vol. 106, no. 2-3, pp. 162–182, 2007.

[41] M. Deshmukh, U. Bhosale *et al.*, "Image fusion and image quality assessment of fused images," *International Journal of Image Processing*, vol. 4, no. 5, p. 484, 2010.

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

[43] G. Qu, D. Zhang, and P. Yan, "Information measure for performance of image fusion," *Electronics Letters*, vol. 38, no. 7, p. 1, 2002.

[44] H. Chen and P. K. Varshney, "A human perception inspired quality metric for image fusion based on regional information," *Information Fusion*, vol. 8, no. 2, pp. 193–207, 2007.

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

[46] Z. Wang and A. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81–84, 2002.

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv: Learning*, 2014.

[48] H. Zhou, W. Wu, Y. Zhang, J. Ma, and H. Ling, "Semantic-supervised infrared and visible image fusion via a dual-discriminator generative adversarial network," *IEEE Transactions on Multimedia*, 2021.

[49] W. Zhao, S. Xie, F. Zhao, Y. He, and H. Lu, "Metafusion: Infrared and visible image fusion via meta-feature embedding from object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 955–13 965.

[50] H. Li, T. Xu, X.-J. Wu, J. Lu, and J. Kittler, "Lrrnet: A novel representation learning guided fusion network for infrared and visible images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.