This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

STFNet: Self-Supervised Transformer for Infrared and Visible Image Fusion

Qiao Liu, Jiatian Pi, Peng Gao, *Member, IEEE*, and Di Yuan, *Member, IEEE*

***Abstract*—Most of the existing infrared and visible image fusion algorithms rely on hand-designed or simple convolution-based fusion strategies. However, these methods cannot explicitly model the contextual relationships between infrared and visible images, thereby limiting their robustness. To this end, we propose a novel Transformer-based feature fusion network for robust image fusion that can explicitly model the contextual relationship between the two modalities. Specifically, our fusion network consists of a detail self-attention module to capture the detail information of each modality and a saliency cross attention module to model contextual relationships between the two modalities. Since these two attention modules can obtain the pixel-level global dependencies, the fusion network has a powerful detail representation ability which is critical to the pixel-level image generation task. Moreover, we propose a deformable convolution-based feature align network to address the slight misaligned problem of the source image pairs, which is beneficial for reducing artifacts. Since there is no ground-truth for the infrared and visible image fusion task, it is essential to train the proposed method in a self-supervised manner. Therefore, we design a self-supervised multi-task loss which contains a structure similarity loss, a frequency consistency loss, and a Fourier spectral consistency loss to train the proposed algorithm. Extensive experimental results on four image fusion benchmarks show that our algorithm obtains competitive performance compared to state-of-the-art algorithms.**

***Index Terms*—Self-supervised, transformer, image fusion, deformable convolution.**

I. Introduction

INFRARED and visible image fusion [1] is a fundamental problem in the field of image processing. The objective of this task is to extract salient visual features from aligned infrared and visible images of a scene and fuse them into a new image that is more perceptually pleasing to the human visual system. Infrared and visible images usually exhibit complementary visual characteristics due to their different imaging principles. For instance, infrared images tend to be clearer than visible images in situations with poor visibility, while visible images are typically richer in texture information than infrared images. By fusing the respective advantages of each modality, image fusion techniques can produce images that are more appealing to the viewer. Therefore, it has various practical applications, such as reconnaissance, border defense, rescue, and video surveillance.

In the past decade, the progress made in infrared and visible image fusion has been significant. One of the main reasons for this progress is the development of more complex and flexible fusion strategies. Early image fusion strategies are usually designed at pixel-level by hand. For example, multi-scale transform-based methods, such as MDBF [2], GFIF [3], ADFA [4], MDLatLRR [5], and MSTN [6] commonly use a fixed coefficient combination method, including the maximum and weighted average etc. Similarly, most of sparse representation-based fusion methods, such as SOMP [7], SRIF [8], and JPCD [9], also adapt these fusion strategies. Some saliency-based fusion methods, such as NSST [10], TSSD [11], and MMGD [12], utilize the information of the saliency map to compute fused coefficients adaptively. Several other multi-scale transform-based methods [13], [14], [15] get the fused coefficients by an optimization method. Although these methods achieve good results, because the hand-designed fusion strategies do not consider the contextual relationships between multiple modalities, which results in the fused image is unfriendly to human perception.

In recent years, some methods attempt to use deep convolution neural networks (CNNs) to improve the effect of infrared and visible image fusion. Most of these methods use CNNs to extract deep features of infrared and visible images and then combine them in a specific feature-level fusion strategy. For example, IFCNN [16] uses two shallow convolution layers with a simple elementwise fusion strategy for end-to-end training. DenseFuse [17] and DLF [18] use a pre-trained dense encoder to get the deep features and then use a l_1-norm based activity level measurement strategy to fuse them. FusionGAN [19], FusionDN [20], DDcGAN [21], GANMcC [22], and RXDNFuse [23] directly use a CNN to perform feature extraction and fusion on two stacked images end-to-end. Similarly, VIFNet [24], STDFusionNet [25], PIAFusion [26], and PMGI [27]

Manuscript received 28 August 2023; accepted 4 October 2023. This work was supported in part by the National Natural Science Foundation of China under Grants 62302073 and 62202362, in part by the Natural Science Foundation of Chongqing under Grants CSTB2022NSCQ-MSX0645 and CSTB2022NSCQ-LZX0040, in part by the Science and Technology Research Program of Chongqing Municipal Education Commission under Grants KJZD-K202200501, KJZD-K20211480, and KJZD-K202114801, in part by the Foundation of National Center for Applied Mathematics in Chongqing under Grant ncamc2022-msxm03, in part by China Postdoctoral Science Foundation under Grant 2022TQ0247, and in part by the Key Project of the Chongqing Technological Innovation and Applications Development Special Program under Grant cstc2021jscx-jbgsX0001. *(Qiao Liu and Jiatian Pi contributed equally to this work.) (Corresponding author: Di Yuan.)*

Qiao Liu and Jiatian Pi are with the National Center for Applied Mathematics, Chongqing Normal University, Chongqing 401331, China (e-mail: liuqiao@stu.hit.edu.cn; pijiatian@cqnu.edu.cn).

Peng Gao is with the School of Cyber Science and Engineering, Qufu Normal University, Qufu 273165, China (e-mail: pgao@qfnu.edu.cn).

Di Yuan is with the Guangzhou Institute of Technology, Xidian University, Guangzhou 511055, China (e-mail: dyuanhit@gmail.com).

Digital Object Identifier 10.1109/TETCI.2024.3352490

2471-285X © 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: University Town Library of Shenzhen. Downloaded on March 27,2024 at 04:02:13 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

fuse the deep features by a concat in channel direction. Furtherly, RFN-Nest [28] designs a residual fusion network to fuse multi-scale features. CSBR [29] use a classification saliency evaluation to get a pixel-level feature weight for fusion. NestFuse [30], AttentionFGAN [31], and MLBF [32] first use a channel and spatial attention mechanism to re-weight feature of each modality and then fuse them by a weighted sum or a standard CNN. Compared with pixel-level hand-designed fusion strategies, these feature-level fusion methods have stronger adaptive ability. However, the core idea of all these feature fusion strategies lie in weighted or convolution. They still do not explicitly consider the contextual relationships between multiple modalities.

To address above-mentioned problem, we propose a Self-supervised Transformer based feature Fusion Network (STFNet) to model the contextual relationships between infrared and visible images for robust image fusion. Different from existing works, the proposed method by modeling the contextual relationships to adaptively capture and fuse the saliency feature of each modality. Specifically, the proposed method contains two modules, feature align network and feature fusion network. Since the features of different modalities are not spatial aligned usually, we design a deformable convolution based alignment network to align these features before fusing them. This module can eliminate artifacts caused by poorly registered of infrared and visible images. After getting the aligned features, we design a Transformer based feature fusion network to get the contextual relationships of these features. This module consists of a Detail Self-Attention (DSA) and a Saliency Cross-Attention (SCA). DSA uses the multi-head self-attention mechanism to capture the critical details of each modality, while SCA explores the multi-head cross-attention to enhance salient features between multiple modalities. These two attention mechanisms can effectively obtain the contextual relationships by modeling the global dependencies of features. In addition to the network architecture, a self-supervised multi-task loss has been designed, which comprises a structure similarity loss, a frequency consistency loss, and a Fourier spectral consistency loss. The structural similarity loss constrains the consistency of brightness, contrast, and structure between the fused image and the source images, while the frequency consistency and Fourier spectral consistency losses ensure that the fused image maintains more details. The fusion network is trained end-to-end using this self-supervised multi-task loss on several multi-modality datasets. The proposed method is validated on four benchmarks, and extensive experimental results show that it achieves competitive performance.

The contributions of this paper are summarized as following three-fold:

- We design an one-stage self-supervised Transformer based fusion network which contains a feature align module and a feature fusion module for robust image fusion. The feature align module solves the misaligned problem and reduces aircrafts in the fused image. The feature fusion module captures and enhances the detail and salient features adaptively by modeling the contextual relationships between multiple modalities effectively.
- We propose a self-supervised multi-task loss which consists of a structure similarity loss, a frequency consistency loss, and a Fourier spectral consistency loss for end-to-end training. The proposed frequency and Fourier spectral losses constrain the detail consistency at the pixel-level and the global-level, respectively, which can preserve more details from source images.
- We conduct extensive experiments on four image fusion benchmarks to verify the effectiveness of the proposed fusion algorithm. The experimental results show that our algorithm achieves competitive performance compared with the state-of-the-art methods.

The rest of this paper is organized as follows. First, we provide an introduction to related infrared and visible image fusion methods and existing Transformer networks in Section II. Then, we describe the main modules and training process of our proposed method in Section III. Next, we present our extensive experimental results, where we evaluate and analyze the effectiveness of the proposed fusion method. Finally, we draw a brief conclusion in Section V.

II. Related Work

A. Deep Learning Based Image Fusion Methods

Due to the powerful representation and adaptive capabilities of deep networks, significant progress has been made in deep network-based infrared and visible image fusion methods in recent years. Deep network-based image fusion methods can be broadly categorized into two groups based on their training methods: two-stage methods and one-stage methods.

Two-stage fusion method: These methods usually first train a generic image reconstruction encoder to extract deep features, and then design a hand-crafted fusion strategy or train a fusion network to get the fusion result. For instance, Li et al. [17] first train an image reconstruct encoder on visible images using a dense net architecture to extract deep feature of both visible and infrared image. Then, they design a l_1-norm based hand-crafted fusion method to get final fusion image. Considering the poor adaptability of the hand-crafted fusion strategy, they then design an attention-based parameter-free fusion method [30] based on the pre-trained reconstruct network. To further improve adaptability of fusion method, they train a residual fusion network [28] based on fixed pre-trained reconstruct network end-to-end. Similarly, Wang et al. [33] first train a Swin Transformer [34] based reconstruct encoder for more robust feature extraction, and then use an extended l_1-norm based hand-crafted fusion strategy. Different from the above methods, DIDFuse [35] decomposes the encoder into two complementary features of details and background to learn reconstruction, and then uses several hand-crafted fusion strategies for fusion. SFAFuse [36] designs a feature adaption reconstruct encoder to obtain more effective feature representation for both visible and infrared images, and then trains an attention based enhancement fusion module. TransFuse [37] proposes a Transformer-based global feature and CNN-based local feature fusion module to reconstruct image, and then also uses a simple hand-crafted fusion strategy. Although these two-stage based fusion methods achieve good performance, the feature learning and fusion process are separated,

Authorized licensed use limited to: University Town Library of Shenzhen. Downloaded on March 27,2024 at 04:02:13 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

which is not conducive to making full use of the complementary relationship between different modalities.

One-stage fusion method: Unlike above-mentioned two-stage fusion approaches, the one-stage fusion method learns the feature encoder and fusion strategy end-to-end simultaneously. For example, FusionGAN [19] first formulates the image fusion task as a generative adversarial problem, and then design an adversarial training manner based fusion method with an adversarial loss and a self-supervised content loss. To improve the fidelity of the fused image, they also propose a dual-discriminator based generative adversarial network for more robust fusion [21]. Considering the detail information is critical for fused result, Ma et al. [38] propose a detail preserving loss on the generative adversarial framework to get more boundaries and textures of the source images. Similarly, Li et al. [31] propose a multi-scale attention module and then integrate it into the generative adversarial framework to adaptively focus on the foreground of infrared image and background details of visible image. Unlike above-mentioned adversarial based fusion method, Long et al. [23] propose a unsupervised end-to-end residual dense network with a feature-level and an image-level similarity constrains to get fused image directly. U2Fusion [39] designs an unified and unsupervised dense network with a data-driven weight strategy of loss functions for fusion tasks. Both DATFuse [40] and CGTF [41] use a convolution and Transformer layer to get more powerful feature. SwinFusion [42] uses a Transformer based fusion strategy with an end-to-end deep encoder. AFT [43] stacks multiple Transformer layers in feature extraction and feature fusion stages simultaneously. In addition, there are several works integrating image registration and fusion into a single framework for unaligned image fusion. For instance, RFNet [44] designs a coarse-to-fine registration network and an attention based fusion method and then combine them into an unified framework. Similarly, Wang et al. [45] integrate a generation based registration network and an corssmodality interaction fusion method for unaligned image fusion. Since these one-stage methods avoid hand-designed fusion rules and separation from feature encoder, they are more robust to image fusion in different scenes. Following this advantage, our method uses an end-to-end self-supervised framework to learn the encoder and fusion strategy simultaneously. However, different from these methods, our method explicitly models the multimodal fusion process using a feature align module and a Transformer based feature fusion module.

B. Visual Transformer

The success of the Transformer [46] model in natural language processing has inspired researchers to explore its applicability in computer vision tasks. In recent years, significant progress has been made in extending the Transformer model to various computer vision applications. Since the first visual Transformer: ViT [47] makes a breakthrough in the classification task, it has been widely applied to different visual tasks, including object detection [48], tracking [49], [50], and segmentation [51] etc. Unlike the convolution operation, which can only focus on local regions of an image, Transfomer can model the long-range dependencies of an image. This feature comes from the attention mechanism and is crucial for many high-level image understanding tasks. For example, DETR [48] uses a self-attention based Transformer to model the relationship between each object and its background for end-to-end object detection. TransT [52] uses both the self-attention and cross-attention based Transformer to fuse the feature of target template and search region for robust visual tracking. SOTR [51] designs a position-aware twin self-attention based Transformer to model the global pixel-level dependency for object segmentation. Different from these methods, our method uses Transformer to model contextual relationships between infrared and visible modality.

C. Self-Supervised Learning

In recent years, self-supervised learning has emerged as a popular technique for solving various visual tasks, including image classification [53] object detection [54], visual tracking [55], etc. Self-supervised learning offers a promising alternative to costly and time-consuming data annotation. It can be broadly categorized into three paradigms based on its role in the target task. First, self-supervised learning pre-trains a general representation learning network using a pretext task on large-scale unlabeled datasets, and then fine-tunes it on downstream tasks. For example, SiamCLR [56] uses an image augmentation based contrastive learning pre-text task to learn a general visual representation, and then fine-tunes it on the part of ImageNet for image classification. Second, self-supervised learning is usually trained as an auxiliary task together with the target task. For example, BF3S [57] uses a self-supervised rotation as an auxiliary task to obtain richer visual representations for classification. Different from the above two paradigms, the third way of self-supervised learning is directly related to the target task. This kind of way usually occurs when the target task cannot provide label information, e.g., image generation task. In this paper, we propose three kind of self-supervised losses for the image fusion task, which belong to the third paradigm.

III. Proposed Fusion Method

A. Algorithm Review

As shown in Fig. 1, our proposed algorithm contains three components, including an encoder-decoder module, a feature align network, and a feature fusion network. What's more, we can see that the proposed framework is an one-stage and can be trained by a self-supervised loss function end-to-end. Given a pair of infrared and visible images $\boldsymbol{I}_r, \boldsymbol{I}_v$, they first through two parameter shared encoders to get the feature representation, respectively. Here, we use a five convolution blocks as the encoder. Since a pair of infrared and visible image are not always strictly aligned, especially the training dataset come from different acquisition devices, we adopt a deformable convolutional network to align the features of the two modalities. This module has a significant effect on dealing with the subtle misalignment issue. After getting the aligned features, we use a Transformer based feature fusion network to model the contextual relationships between the infrared and visible images. This network

Authorized licensed use limited to: University Town Library of Shenzhen. Downloaded on March 27,2024 at 04:02:13 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

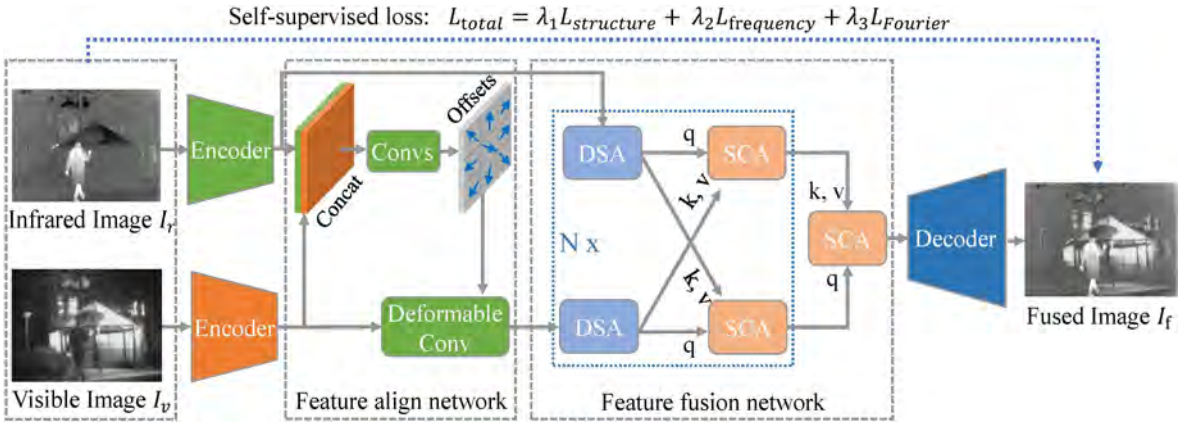

Fig. 1. Framework of our image fusion algorithm, STFNet. It contains an encoder-decoder module, a feature align network, and a feature fusion network. The feature fusion network mainly contains two components: DSA and SCA (see Fig. 3).

adaptively obtains salient features of two modalities and fuses them through a multi-head self-attention and a multi-head cross-attention modules. Once the fused features are obtained, we use a deconvolution network as decoder to restore the fused image $\boldsymbol{I}_f$. Since shallow convolution feature has a significant impact on image restoration tasks, we use a feature concatenation operation between encoder and decoder just like UNet [58].

B. Encoder-Decoder

We use an encoder-decoder architecture as the framework of our fusion algorithm. The proposed feature align network and feature fusion network can be easy plugged into this framework and can be trained end-to-end. The encoder consists of five convolution blocks, which is used to obtain the deep convolution features of infrared and visible images. Each convolution block has two convolution layers (kernel size $= 3 \times 3$, stride $= 1$, padding $= 1$) and a maxpooling layer (kernel size $= 2$) except for the last block. Each convolution layer is followed by an ReLU activation unit. Given a pair of infrared and visible input image $\boldsymbol{I}_r \in \mathbb{R}^{H_{im} \times W_{im} \times 1}$, $\boldsymbol{I}_v \in \mathbb{R}^{H_{im} \times W_{im} \times 1}$, the output of the encoder are two corresponding feature maps of each convolution block with different resolution. Before fusing these features, we first register these features using the proposed feature align network on each convolution block. Then we use the feature fusion network to get the fused feature map on the last convolution block, since the Transformer-based feature fusion needs to consume high computing resources on the feature map with large resolution.

The decoder receives the fused feature and concatenates the aligned feature of each convolution block to restore the fused image. The decoder consists of four convolution blocks and each block contains a deconvolution layer (kernel size $= 4 \times 4$, stride $= 2$, padding $= 1$) and two convolution layers (kernel size $= 3 \times 3$, stride $= 1$, padding $= 1$). For each block, we first use a deconvolution to upsample the fused feature map, and then concatenate it with the corresponding convolution features of the encoder in channel direction. Finally, we feed the concatenated feature into two convolution layers to restore the fused image gradually.

(a) Image pair of the KAIST [62] dataset

(b) Image pair of the M3FD [63] dataset

Fig. 2. Examples of the slightly misaligned infrared and visible image pairs on two multi-modality datasets.

C. Feature Align Network

For the image fusion task, the source image pairs are usually assumed to be strictly aligned by default. Currently, most existed multi-modality fusion datasets usually use image registration algorithms [59], [60] to align infrared and visible images due to the expensive equipment to capture strictly aligned images. However, these registration methods cannot completely and strictly align all images of the two modalities. This results in a lot of slightly misaligned image pairs in the dataset, as shown in Fig. 2. This phenomenon causes some image fusion algorithms to introduce artifacts into the fused image.

Different from previous unaligned image fusion methods, e.g., RFNet [44] and CGRP [45], which assume that all image pairs with a large misalignment and are not preprocessed by an image registration algorithm, we assume that there just exists silghtly misaligned of image pairs in the fusion dataset, which have preprocessed by an image registration algorithm.

Authorized licensed use limited to: University Town Library of Shenzhen. Downloaded on March 27,2024 at 04:02:13 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Therefore, we suggest that it does not need a complex image registration module instead of using a feature alignment model is more suitable for this situation.

To solve this problem, we design a deformable convolution based network to align the infrared and visible images in the feature space. We first extract deep convolution features of a pair of infrared and visible images by the encoder. Then, we concat these two features in channel direction and use several convolutions to predict a series of offsets for each feature point. These offsets represent the offsets of each feature point in the X-axis and Y-axis directions in 2D space. Finally, we use a deformable convolution [61] to align the visible image to the infrared image. Given a pair of infrared and visible image's feature maps $\boldsymbol{F}_{ir} \in \mathbb{R}^{H\times W\times C}$ and $\boldsymbol{F}_{vis} \in \mathbb{R}^{H\times W\times C}$, the offset $\boldsymbol{O} \in \mathbb{R}^{H\times W\times 2^{\cdot}K}$ can be predicted by:

$$\boldsymbol{O} = Convs(Concat(\boldsymbol{F}_{ir}, \boldsymbol{F}_{vis})), \tag{1}$$

where $Concat(,)$ and $Convs()$ denote the feature concatenation in channel direction and two $k \times k$ convolution kernels, respectively. The channel number K in offset equals $k * k$. After getting the offset, we can use a deformable convolution to get the new aligned visible image's feature:

$$\boldsymbol{F}_{vis\_aligned} = DConv(\boldsymbol{F}_{vis}, \boldsymbol{O}), \tag{2}$$

where $DConv(,)$ denotes the deformable convolution [61] operation, which uses a linear interpolation to apply the offset to the feature of the visible image.

D. Feature Fusion Network

How to fuse the features of infrared and visible images is crucial for the image fusion task. As we known that learning saliency features of a single modality is related to its surrounding context, while the fusion of saliency features of two modalities is related to the context of both modalities. From the previous introduction, we know that most of the existing feature fusion strategies use a simple convolution or weighted average method. However, these strategies do not explicitly consider the contextual relationships between infrared and visible images.

To model the contextual relationship, we propose a Transformer based feature fusion network, as shown Fig. 1. The proposed feature fusion network mainly contains a detail self-attention (DSA) module and a saliency cross-attention (SCA) module. These two modules model the contextual relationships of a single modality and the contextual relationships between two modalities through a multi-head self-attention mechanism and a multi-head cross-attention mechanism, respectively. Specifically, we first use two parallel DSA and SCA to form a feature fusion unit, then cascade N identical units, and finally use a SCA to select the final fused feature for reconstruction, which is similar to SwinFusion [42]. However, different from SwinFusion which divides the feature of an image into a series of non-overlapping local windows using a shifted window mechanism, and then model the relationship of them using the multi-head attention mechanism. We do not divide feature windows but model the contextual relationship on all feature points. We suggest that our method can more fully model the contextual relationship of multimodal image pairs than SwinFusion. At following, we mainly describe the structure of the proposed DSA and SCA modules.

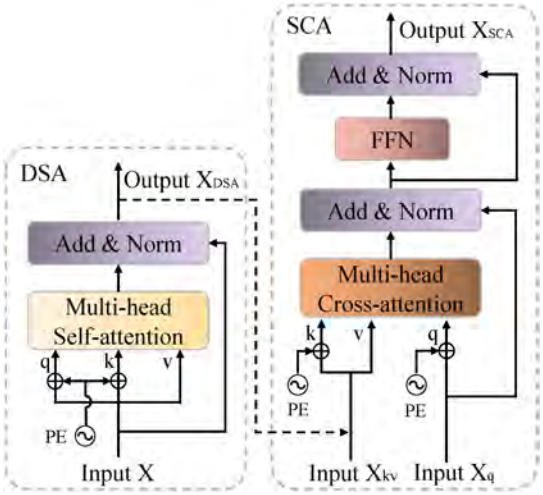

Fig. 3. Structure of the detail self-attention (DSA) and saliency cross-attention (SCA) modules.

Detail self-attention (DSA): As shown in the left of Fig. 3, the multi-head self-attention is the central component of the DSA module, comprising several attention units. Each attention unit takes in queries $\boldsymbol{Q}$, keys $\boldsymbol{K}$ and values $\boldsymbol{V}$ as input. Specifically, a single attention unit can be defined as follows:

$$Attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathsf{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}, \tag{3}$$

where d_k is the key dimensionality. Unlike the aforementioned attention models, the multi-head attention approach employs multiple parallel attention modules to extract valuable information from various viewpoints. This technique has been employed in our study to enable the feature network to focus more closely on the details of distinct regions. The mathematical formulation of multi-head attention is as follows:

$$\begin{aligned} MHA(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) &= Concat(\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_n)\boldsymbol{W}, \\ \boldsymbol{A}_i &= Attention(\boldsymbol{Q}\boldsymbol{W}_i^Q, \boldsymbol{K}\boldsymbol{W}_i^K, \boldsymbol{V}\boldsymbol{W}_i^V), \end{aligned} \tag{4}$$

where $\boldsymbol{W} \in \mathbb{R}^{d_m\times nd_v}$, $\boldsymbol{W}_i^Q \in \mathbb{R}^{d_m\times d_k}$, $\boldsymbol{W}_i^K \in \mathbb{R}^{d_m\times d_k}$, $\boldsymbol{W}_i^V \in \mathbb{R}^{d_m\times d_v}$ are parameter matrices, and n denote the number of attention unit in the multi-head attention. Here, we set $n = 8$, $d_m = 256$, and $d_k = d_v = d_m/n = 32$.

Given an image's feature map $\boldsymbol{F} \in \mathbb{R}^{H\times W\times C}$, since DSA requires a serialized feature vector as input, we first reshape and reduce dimension of it to form a new feature $\boldsymbol{X} \in \mathbb{R}^{HW\times d}$, the output feature $\boldsymbol{X}_{DSA} \in \mathbb{R}^{HW\times d}$ can be formulated as:

$$\boldsymbol{X}_{DSA} = \boldsymbol{X} + MHA(\boldsymbol{X} + \boldsymbol{P}, \boldsymbol{F} + \boldsymbol{P}, \boldsymbol{X}), \tag{5}$$

where $\boldsymbol{P} \in \mathbb{R}^{HW\times d}$ represents the positional encoding using a sine function like DERT [48].

Authorized licensed use limited to: University Town Library of Shenzhen. Downloaded on March 27,2024 at 04:02:13 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Saliency cross-attention (SCA): As shown in the right of Fig. 3, the core component of SCA is the multi-head cross-attention. Similar to DSA, the multi-head cross-attention also contains multiple attention units. However, different from DSA, this attention needs two different inputs. Since the input feature vector does not contain the spatial information, we also use a positional encoding like DSA. What's more, we use a feed forward neural network (FFN) to enhance the fitting ability, which uses two linear layers and a ReLU activation unit as shown in below:

$$FFN(\boldsymbol{x}) = \max(0, \boldsymbol{x}\boldsymbol{W}_1 + \boldsymbol{b}_1)\boldsymbol{W}_2 + \boldsymbol{b}_2, \tag{6}$$

where $\boldsymbol{W}_1, \boldsymbol{W}_2$ and $\boldsymbol{b}_1, \boldsymbol{b}_2$ denotes the weight matrices and bias vectors, respectively. Given two input features $\boldsymbol{X}_q \in \mathbb{R}^{HW\times d}$ and $\boldsymbol{X}_{kv} \in \mathbb{R}^{HW\times d}$, the output feature $\boldsymbol{X}_{SCA} \in \mathbb{R}^{HW\times d}$ can be formulated as:

$$\begin{aligned} \boldsymbol{X}_{SCA} &= \boldsymbol{X}_{CA} + FFN(\boldsymbol{X}_{CA}), \\ \boldsymbol{X}_{CA} &= \boldsymbol{X}_q + MHA(\boldsymbol{X}_q + \boldsymbol{P}_q, \boldsymbol{X}_{kv} + \boldsymbol{P}_{kv}, \boldsymbol{X}_{kv}), \end{aligned} \tag{7}$$

where $\boldsymbol{P}_q \in \mathbb{R}^{HW\times d}$ and $\boldsymbol{P}_{kv} \in \mathbb{R}^{HW\times d}$ denote the corresponding positional encoding with the input $\boldsymbol{X}_q$ and $\boldsymbol{X}_{kv}$, respectively.

E. Self-Supervised Multi-Task Loss

Structure similarity loss: Different from most visual tasks, image fusion task has not groundtruth. Therefore, it is critical to design a proper self-supervised loss function to train the fusion model. Following many previous works [30], [39], [42], we employ a structural similarity loss which ensures that the fused image remains consistent with the light, contrast, and structural features of the source images. The structure similarity loss can be described as:

$$\begin{aligned} \mathcal{L}_{structure} = {} & w_1(1 - MSSSIM(\boldsymbol{I}_f, \boldsymbol{I}_v)) \\ & + w_2(1 - MSSSIM(\boldsymbol{I}_f, \boldsymbol{I}_r)), \end{aligned} \tag{8}$$

where $MSSSIM(,)$ denotes the multi-scale structure similarity [64], and w_1, w_2 represent the balance parameters. Here, we consider the contribution of these two balance terms are the same, i.e., $w_1 = w_2 = 0.5$.

Frequency consistency loss: In addition to the structure similarity, we also hope the detail and intensity of fused image to be similar to the source images. We know that high-frequency components of an image contain its detail information, while the low-frequency components convey its intensity information. Therefore, we first use a Laplace operator to decompose the source image into a high-frequency image and a low-frequency image. Then, a constraint based on frequency consistency is proposed as follows:

$$\begin{aligned} \mathcal{L}_{frequency} = {} & \|La(\boldsymbol{I}_f) - \mathrm{Max}(La(\boldsymbol{I}_v), La(\boldsymbol{I}_r)))\|_1 \\ & + \|(\boldsymbol{I}_f - La(\boldsymbol{I}_f)) - Max(\boldsymbol{I}_v - La(\boldsymbol{I}_v), \boldsymbol{I}_r \\ & - La(\boldsymbol{I}_r))\|_1, \end{aligned} \tag{9}$$

where $La(\cdot)$ denotes a Laplace operator and $Max(,)$ denotes the element-wise maximum operation.

Fourier spectral consistency loss: Furthermore, different from the above-proposed frequency loss which constrains the detail consistency in the pixel-level, we also expect the details of the fused image to be consistent with the source image in the global-level. To this end, we transform the image into the Fourier domain and require the high frequency part of the fused image to be consistent with the source images. Since the high-frequency part of an image in the Fourier domain is calculated from all the pixels in the image, it can constrain the fused image to retain more details from the global-level. Specifically, we first transform an image $\boldsymbol{I}$ into the Fourier spectral space using Discrete Fourier Transform $\mathcal{F}$:

$$\begin{aligned} \mathcal{F}(\boldsymbol{I})(x,y) = {} & \frac{1}{H_{im}W_{im}} \sum_{h=0}^{H_{im}-1} \sum_{w=0}^{W_{im}-1} \\ & e^{-2\pi i \cdot \frac{ha}{H_{im}}} e^{-2\pi i \cdot \frac{wb}{W_{im}}} \cdot \boldsymbol{I}(h,w), \end{aligned} \tag{10}$$

where $x = 0, 1, \ldots, H_{im} - 1$ and $y = 0, 1, \ldots, W_{im} - 1$. To easy train the model, then we convert $\mathcal{F}$ from the complex number doamin to the real number domain:

$$\begin{aligned} \mathcal{F}^R(\boldsymbol{I})(x,y) = {} & \log\Big(1 + \sqrt{[Re(\mathcal{F}(\boldsymbol{I})(x,y))]^2} \\ & + \sqrt{[Im(\mathcal{F}(\boldsymbol{I})(x,y))]^2} + \epsilon\Big), \end{aligned} \tag{11}$$

where $Re(\cdot)$, $Im(\cdot)$ are the real part and imaginary part of $\mathcal{F}(\boldsymbol{I})(x,y))$ respectively. Based on these operations, we propose a high-frequency of Fourier spectral consistency loss as following:

$$\begin{aligned} \mathcal{L}_{Fourier} = {} & w_1(\|\mathcal{F}_H^R(\boldsymbol{I}_f) - \mathcal{F}_H^R(\boldsymbol{I}_v)\|_1) \\ & + w_2(\|\mathcal{F}_H^R(\boldsymbol{I}_f) - \mathcal{F}_H^R(\boldsymbol{I}_r)\|_1), \\ \mathcal{F}_H^R(\boldsymbol{x}) = {} & \mathcal{F}^R(\boldsymbol{x}) \cdot \mathcal{M}_H, \end{aligned} \tag{12}$$

where $\mathcal{M}_H$ is a circle mask which gets the high frequency signals from the overall Fourier spectral space.

Finally, we use above-mentioned three losses to form a new multi-task self-supervised loss to train the proposed fusion network end-to-end:

$$\mathcal{L}_{total} = \lambda_1\mathcal{L}_{structure} + \lambda_2\mathcal{L}_{frequency} + \lambda_3\mathcal{L}_{Fourier}, \tag{13}$$

where $\lambda_1, \lambda_2, \lambda_3$ are the weight of each loss.

IV. Experiment

A. Implementation Details

Experiment settings: We train our fusion network on several multi-modality datasets, including KAIST [62], LLVIP [65], M3FD [63], MSRS [26], and VLIRVDIF [66]. These datasets contain close to $200\tilde{}K$ image pairs in total. At the training stage, we resize all of the image pairs to 256×256 and normalize them to $[-1, 1]$. We set convolution kernel size $k = 3$ in the feature align network, the number of feature fusion unit $N = 4$ in the feature fusion network, and the radius of the mask $\mathcal{M}_H$ is 21. We train our fusion network 10 epochs using Adam optimizer with an exponentially decaying learning rate of 0.0001 and $batchsize = 8$. We set the weight of each loss $\lambda_1 = 1, \lambda_2 = 10$,

Authorized licensed use limited to: University Town Library of Shenzhen. Downloaded on March 27,2024 at 04:02:13 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE I
ABLATION STUDIES OF THE NETWORK ARCHITECTURE OF OUR IMAGE FUSION ALGORITHM ON THE TNO AND ROADSCENE DATASETS

Method				TNO [69]					RoadScene [39]				
Base	DSA	SCA	FAN	MI ↑	$Q^{AB/F}$ ↑	$N^{AB/F}$ ↓	VIF ↑	SSIM ↑	MI	$Q^{AB/F}$	$N^{AB/F}$	VIF	SSIM
✓				3.9546	0.5088	0.0371	0.9172	0.8821	4.6195	0.4528	0.0492	0.8782	**0.7916**
✓	✓	✓		4.2149	0.5208	0.0360	0.9375	**0.8873**	4.7760	0.4553	0.0464	0.8865	0.7911
✓	✓	✓	✓	**4.7786**	**0.5373**	**0.0350**	**0.9707**	0.8579	**5.4000**	**0.4613**	**0.0398**	**0.8926**	0.7714
✓	✓		✓	4.0632	0.4649	0.0403	0.9548	0.8459	4.8982	0.4342	0.0443	0.8895	0.7605
✓		✓	✓	4.0026	0.4693	0.0385	0.9458	0.8499	4.8615	0.4432	0.0463	0.8868	0.7608

Base, DSA, SCA, and fan denote the baseline method, detail sefl-attention, saliency cross-attention, and feature align network respectively. The bold represents the best score.

and $\lambda_3 = 1$. The proposed method achieves an average speed of 4.7 and 11.7 frames per second on the TNO and RoadScene datasets, respectively. All of experiments are conducted on a PC with a NVIDIA RTX A4000 GPU with PyTorch framework.

Evaluation datasets: We use four image fusion datasets for evaluation, including TNO [67], Roadsense [39], MSRS [26], and LLVIP [65]. Unlike some previous methods that only select a subset of image pairs for evaluation, to facilitate fair comparison, we use all 42 image pairs from the TNO dataset for evaluation. However, TNO is an older dataset with a few scenes and most images are of low resolution. Different from TNO, Roadsense [39] and MSRS [26] mainly focuses on the road scenario and have 221 and 361 image pairs respectively, with higher resolution. These two datasets captured from vehicle-mounted and hand-held cameras within daytime and nighttime environment. LLVIP [65] mainly contains 3463 surveillance scenarios image pairs and their resolution is up to 1080×720. All of image pairs of this dataset are captured from a surveillance camera in a low-light condition.

Evaluation metrics: We use five kind of metrics for quantitative evaluation [68]. The first type is the metric based on information entropy, which includes Entropy(EN), Mutual Information (MI), and Peak Signal-to-Noise Ratio (PSNR). The second type is the metric based on image feature, which includes Spatial Frequency (SF), Standard Deviation (SD), Gradient-based fusion performance ($Q^{AB/F}$), and Artifact based fusion performance ($N^{AB/F}$). The third type is the metric based on image structure, e.g., Structural Similarity Index Measure (SSIM). The fourth type is the metric based on correlation including Correlation Coefficient (CC), and Sum of Correlation Differences (SCD). The last type is metric based on human perception, e.g., Visual Information Fidelity (VIF).

Comparison methods: We select nine more recently deep learning based image fusion algorithms for comparison. These methods include two-stage based methods, e.g., DenseFuse [17], RFN-Nest [28] and one-stage based method, e.g., FusionGAN [19], U2Fusion [39], IFCNN [16], SDNet [69], PIAFusion [26], PMGI [27], SeAFusion [70], SwinFusion [42], and DATFuse [40].

B. Ablation Studies

Network architecture: As shown in Table I , we show several group comparison experiments using different components of the proposed network. The baseline method is an encoder-decoder architecture without any other modules. The first two rows of Table I shows that the proposed feature fusion network (DSA+SCA) boosts the MI metric by a large margin on both two datasets. This demonstrate that the feature fusion network is good at capturing the contextual information between infrared and visible images and transferring them. We suggest that this is mainly because the proposed detail self-attention and saliency cross attention are effective to obtain the details of images and contextual dependencies between infrared and visible images. Although the feature fusion network does not show an advantage or even a decline on the SSIM metric, it obtains a remarkable improvement on the VIF metric. This illustrates that the feature fusion network fuses contextual information selectively rather than fuses as much information as possible. On the $Q^{AB/F}$ metric, the feature fusion network also shows an obviously gain on both two datasets, which demonstrates that it is effective. To demonstrate that both the DAS and SCA modules of the feature fusion network can boost the fusion performance, we conduct two experiments without using them, as shown in the last two rows of Table I. We can see that removing any of them will cause a decline in the performance. From the second to third row of Table I, we can see that the feature align network further improves fusion performance on the most metrics. Especially, it achieves a remarkable improvement on the $N^{AB/F}$ metric. This demonstrate that it is effective to introduce fewer artifacts into fused images. To demonstrate the effect of the feature alignment network more intuitively, we show a comparison of two slightly misaligned image pairs with and without the alignment module, as shown in Fig. 4. We can see that the first pair of images has obvious artifacts in the person's leg area when feature align network is not used, and the second pair of images has more significant artifacts on the right edge of the person, while these artifacts disappeared when the feature alignment module is used.

Loss function: We present the effect of the each term of self-supervised multi-task loss to the fusion results, as shown in Table II . From the first two rows, we can see that it has limited performance when we only use the structure similarity loss. When we use the structure similarity loss and frequency loss simultaneously, the performance has a huge boosting on all the metrics. This shows that frequecy information is crucial for the image fusion task. The last two rows of Table II shows that the Fourier spectral consistency loss further improve the fusion performance on the all of these metrics. Especially, on the $Q^{AB/F}$ metric, the Fourier spectral consistency loss obtains about 1% gains on both two datasets. This demonstrate that the Fourier spectral consistency loss can boost the model to persever the high frequency details of the source images effectively.

Authorized licensed use limited to: University Town Library of Shenzhen. Downloaded on March 27,2024 at 04:02:13 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Fig. 4. Visualized comparison of the proposed fusion method with and without using feature align network (FAN) on two slightly misaligned image pairs in RoadScene dataset. The red and green bounding boxes denote the misaligned region of source images and local region of fused images, respectively.

TABLE II
ABLATION STUDIES OF OUR SELF-SUPERVISED MULTI-TASK LOSS ON THE TNO AND ROADSCENE BENCHMARKS

Loss			TNO [69]					RoadScene [39]				
$\mathcal{L}_s$	$\mathcal{L}_f$	$\mathcal{L}_F$	MI ↑	$Q^{AB/F}$ ↑	$N^{AB/F}$ ↓	VIF ↑	SSIM ↑	MI ↑	$Q^{AB/F}$ ↑	$N^{AB/F}$ ↓	VIF ↑	SSIM ↑
✓			2.1418	0.2749	0.0472	0.6152	0.8323	3.6596	0.4131	0.0898	0.7341	**0.8581**
✓	✓		4.6149	0.5214	**0.0298**	0.9425	**0.8683**	5.0779	0.4457	0.0402	0.8838	0.7817
✓	✓	✓	**4.7786**	**0.5373**	0.0350	**0.9707**	0.8579	**5.4000**	**0.4613**	**0.0398**	**0.8926**	0.7714

$\mathcal{L}_s$, $\mathcal{L}_f$, and $\mathcal{L}_F$ denote the structure similarity loss, frequency consistency loss, and fourier spectral consistency loss respectively.

TABLE III
COMPARISON WITH STATE-OF-THE-ART FUSION METHODS ON THE TNO [67] AND ROADSCENE [39] DATASETS

Method \ Metric	TNO [69]					Roadscene [39]				
	EN ↑	MI ↑	$Q^{AB/F}$ ↑	$N^{AB/F}$ ↓	VIF ↑	SD ↑	MI ↑	$Q^{AB/F}$ ↑	$N^{AB/F}$ ↓	VIF ↑
DenseFuse [17]	6.8192	2.3019	0.4457	0.0788	0.8174	9.5851	2.7174	0.3648	0.0541	0.6695
RFN-Nest [28]	6.9631	2.1184	0.3341	0.0649	0.8182	10.0803	2.7476	0.2982	*0.0753*	0.7247
FusionGAN [19]	6.5580	2.3352	0.2340	0.0770	0.6541	9.9856	2.7555	0.2579	0.1078	0.5784
U2Fusion [39]	*6.9966*	2.0102	0.4262	0.3047	0.8196	9.6962	2.6446	0.4710	0.1067	0.6504
IFCNN [16]	6.8539	2.0526	0.4805	0.2733	0.7869	10.1334	2.9152	0.5348	0.1740	0.7336
SDNet [71]	6.6948	2.2605	0.4294	0.2111	0.7591	10.0104	3.2476	*0.5089*	0.2001	0.7679
PIAFusion [26]	6.8142	3.3575	*0.5280*	0.0814	*0.8714*	*10.1406*	*3.6331*	0.4362	0.0924	0.8373
PMGI [27]	7.0180	2.3520	0.4117	0.1130	0.8691	10.0255	3.2712	0.4442	0.1262	0.7934
SeAFusion [72]	**7.1335**	2.8382	0.4871	0.2798	**0.9810**	**10.7929**	3.0287	0.4927	0.2471	*0.8139*
SwinFusion [42]	6.4894	1.8956	**0.5452**	*0.0706*	0.7022	9.6532	2.5839	**0.5875**	0.0916	0.7010
DATFuse [40]	6.4531	*3.1323*	0.4971	0.1178	0.7414	10.0316	3.6679	0.4657	0.0673	0.7269
STFNet (Ours)	6.8081	**4.7786**	0.5373	**0.0350**	0.9707	10.1481	**5.4000**	0.4613	**0.0398**	**0.8926**

The bold, underline, and italic represent the best, the second-best, and the third-best score, respectively.

C. Comparison With State-of-The-Arts

Results on TNO: The left part of Table III shows that our proposed algorithm outperforms other methods, achieving top three scores on most of the metrics and comparable performance on the others. Specifically, our method, STFNet, shows significant improvement on the MI metric, indicating that it can extract more information from the source images. Additionally, STFNet achieves the second-best score (0.5373) on the $Q^{AB/F}$ metric, suggesting that it can extract edge information more effectively. We attribute these good results to the fact that the proposed feature fusion network can better model the contextual relationship between the multimodal images. What's more, our method gets the best score on the $N^{AB/F}$ metric. This shows that the fused image generated by our method introduces fewer artifacts compared with other methods. This is mainly because our method contains a feature align network which can solve the slight misaligned problem of the infrared and visible image pair. Although the proposed method is not achieve the best structure similarity with the source images, it obtains the second-best visual information fidelity. This means that the fused image generated by our method is more friendly for human perception.

Authorized licensed use limited to: University Town Library of Shenzhen. Downloaded on March 27,2024 at 04:02:13 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE IV
COMPARISON WITH STATE-OF-THE-ART FUSION METHODS ON THE MSRS [26] AND LLVIP [65] DATASET

Method \ Metric	MSRS [67]					LLVIP [66]				
	EN ↑	MI ↑	SD ↑	$Q^{AB/F}$ ↑	VIF ↑	EN ↑	MI ↑	SD ↑	$Q^{AB/F}$ ↑	VIF ↑
DenseFuse [17]	5.9308	2.6659	7.4232	0.3653	0.6997	6.8428	2.7118	9.3773	0.3481	0.7253
RFN-Nest [28]	6.1965	2.4587	7.8002	0.3874	0.7372	7.0411	2.5724	9.6094	0.2825	0.7583
FusionGAN [19]	5.4313	1.8934	5.9404	0.1389	0.5007	6.4420	2.8573	8.7402	0.2252	0.4980
U2Fusion [39]	4.9532	1.9584	5.9287	0.3139	0.4396	6.1439	2.4633	8.3305	0.3083	0.5355
IFCNN [16]	6.4394	2.8500	7.9829	0.5348	0.8515	7.2243	2.9689	9.7278	*0.6454*	0.8095
SDNet [71]	5.2450	1.7133	5.7762	0.5089	0.4333	6.8920	3.0276	9.4199	0.5425	0.6607
PIAFusion [26]	6.6373	*3.9909*	8.4057	0.6604	**1.0422**	*7.3453*	3.3911	9.7485	**0.6781**	0.9353
PMGI [27]	6.2427	2.1764	7.8239	0.4121	0.6706	7.0729	*3.4866*	*9.8790*	0.3910	0.7373
SeAFusion [72]	**6.6514**	4.0374	*8.3770*	**0.6623**	0.9859	**7.4180**	3.8572	9.8452	0.6222	**0.9370**
SwinFusion [42]	6.0666	2.2887	7.4458	0.6017	0.7143	6.9369	2.5901	9.3494	0.6746	0.7642
DATFuse [40]	6.4795	3.8965	**8.5079**	*0.6349*	0.8265	7.1110	3.8558	**10.0974**	0.4621	0.7762
STFNet (Ours)	*6.6114*	**4.8442**	8.3624	0.6313	*0.9574*	7.4147	**4.8582**	9.9076	0.4868	*0.9025*

The bold, underline, and italic represent the best, the second-best, and the third-best score, respectively.

To demonstrate the proposed method achieves favorable performance more intuitively, we compare fused images of several state-of-the-art methods with our method on six challenging image pairs, as shown in Fig. 5. The first three columns show that the our method preserves the salient intensity information in the infrared and visible images, while most of other methods weaken these intensity information to some extent. As shown in the last three columns, our method can better transfer edge details and textures from the source image to the fused image while preserving the intensity information well.

Results on RoadScene: The right part of Table III illustrates that our proposed method also achieves top three performance on half of the metrics and competitive results on the others. Notably, our method obtains the best scores on the MI and $N^{AB/F}$ metrics of the RoadScene dataset, which are consistent with the results on the TNO dataset. When compared to the SeAFusion [70] method, our proposed algorithm exhibits an 8% improvement on the VIF metric. Interestingly, this indicates that high-level vision tasks driven fusion methods may not always be beneficial for human perception. Furthermore, our method shows a significant improvement on the MI metric than SwinFusion [42] which also uses a Transfomer fusion strategy. This suggests that our fusion strategy excels at transferring information from both source images rather than maintaining the same content as one of them. To validate our method introduces fewer aircrafts into fused images, we show a visualization comparison, as shown in Fig. 6. We can see that most of the methods produce noticeable artifacts at edges of the tree, while our method is able to produce more natural image without any artifacts. In addition to these two metrics, our method achieves the best score on the VIF metric and the second-best score on the SD metric. These results demonstrate that the fused images generated by our method are more friendly to the human perception. As shown in Fig. 7, most methods can focus on salient objects in the infrared image, however our method can not only transfer saliency information in the infrared image but also transfer details in the visible image (e.g., bicycle wheel). *Results on MSRS:* The left part of Table IV indicates that our proposed method achieves competitive performance on most metrics. For instance, on the MI and $Q^{AB/F}$ metrics, our method obtains the best and fourth-best scores, respectively, demonstrating its robustness in transferring edge information into the fused image in different scenarios. On the EN and VIF metrics, our method achieves the third-best results, indicating its ability to capture details and preserve intensity from source images. When compared hand-crafted fusion strategy based methods, DenseFuse [17] and RFN-Nest [28] which cannot consider contextal information, our method outperforms these two methods on the MI and $Q^{AB/F}$ metric remarkably. This show that the propsoed fusion strategy can model contextal information and transfer them into the fused images effectivley. Compared to the PIAFusion [26] method, which achieves the best score on the VIF metric, our proposed algorithm introduces fewer artifacts while maintaining similar performance on the VIF metric. This demonstrates that our fusion strategy is robust to illumination variations due to its strong adaptive capability. Figs. 8 and 9 present daytime and nighttime visualization comparisons of fused images. Our proposed method effectively transfers salient intensity and detail information from the source images regardless of lighting changes.

Results on LLVIP: The right part of Table IV demonstrates that our proposed method achieves similar results as on the other three datasets. Specifically, the proposed method obtains the second-best score on the EN metric. These results suggest that our method can generate fused images with richer information compared to most other fusion methods. Additionally, our method exhibits favorable performance on the MI and VIF metrics, indicating its good generalization ability to different scenarios. Compared with one-stage method IFCNN [16] which uses a simple convolution based fusion strategy, our method achieves a 10% gain on the VIF metric, despite IFCNN achieves a higher score on the $Q^{AB/F}$. This means that the proposed fusion network pays more attention to global information of the source images, which is crucial to human perception. Figs. 10 and 11 show that the proposed method is good at to transfer the details of low-light visible images and intensity of infrared images. When compared to the one-stage method IFCNN [16],

Authorized licensed use limited to: University Town Library of Shenzhen. Downloaded on March 27,2024 at 04:02:13 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Fig. 5. Visualized comparison of our image fusion algorithm (STFNet) with others on several challenging image pairs of the TNO [67] dataset. From top to bottom, each row represents source infrared image, source visible image, fused images of DenseFuse, FusionGAN, U2Fusion, IFCNN, SDNet, PMGI, SwinFusion, and our STFNet, respectively.

Authorized licensed use limited to: University Town Library of Shenzhen. Downloaded on March 27,2024 at 04:02:13 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Fig. 6. Visualized comparison of the proposed fusion method (STFNet) with nine state-of-the-art methods on the 'FLIR_06832' image pair of the RoadScene [39] dataset.

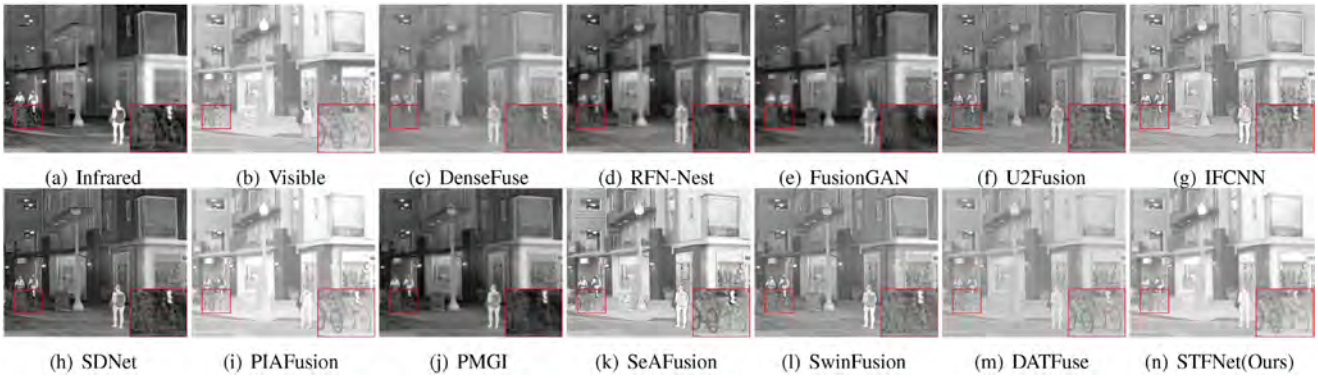

Fig. 7. Visualized comparison of the proposed fusion method (STFNet) with nine state-of-the-art methods on the 'FLIR_08835' image pair of the RoadScene [39] dataset.

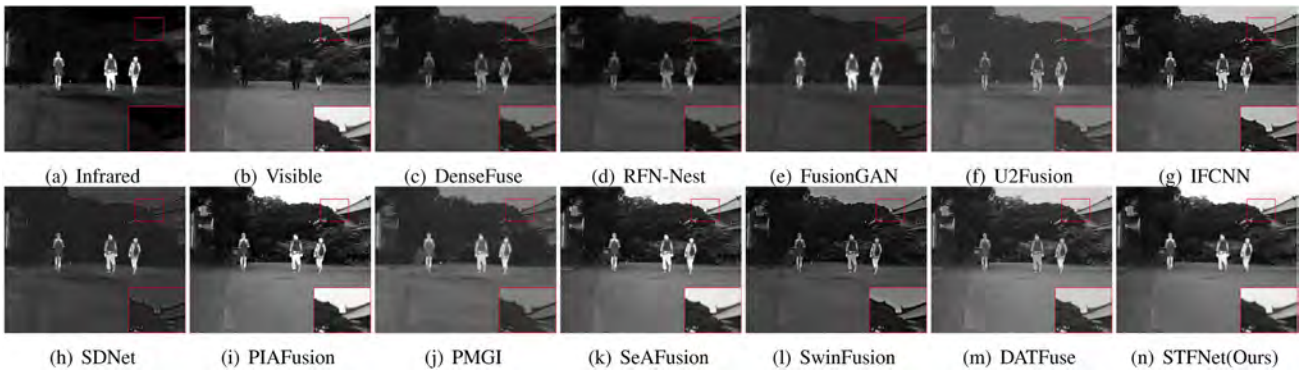

Fig. 8. Visualized comparison of the proposed fusion method (STFNet) with nine state-of-the-art methods on the daytime '00634D' image pair of the MSRS [26] dataset.

which uses a simple convolution-based fusion strategy, our proposed algorithm exhibits an improvement of 10% on the VIF metric, despite IFCNN achieving a higher score on the $Q^{AB/F}$ metric. This suggests that the proposed fusion network places greater emphasis on global information from the source images, which is crucial for human perception. Figs. 10 and 11 demonstrate that our proposed method excels at transferring the details of low-light visible images and the intensity of infrared images. However, when the low-light image contains severe exposure phenomena, the proposed method cannot effectively suppress the exposed regions, which are usually considered as rich details and preserved into the fused image. We suggest that this is mainly because the proposed method does not have high-level semantic supervision and can only reconstruct images from the pixel level. In the future, we will explore high-level semantic supervision to address this challenge.

Authorized licensed use limited to: University Town Library of Shenzhen. Downloaded on March 27,2024 at 04:02:13 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

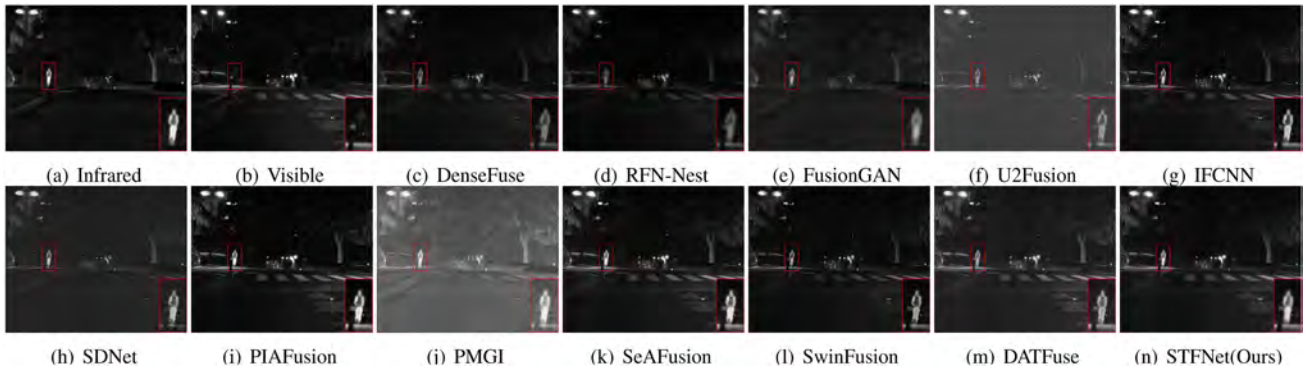

Fig. 9. Visualized comparison of the proposed fusion method (STFNet) with nine state-of-the-art methods on the nighttime '00882 N' image pair of the MSRS [26] dataset.

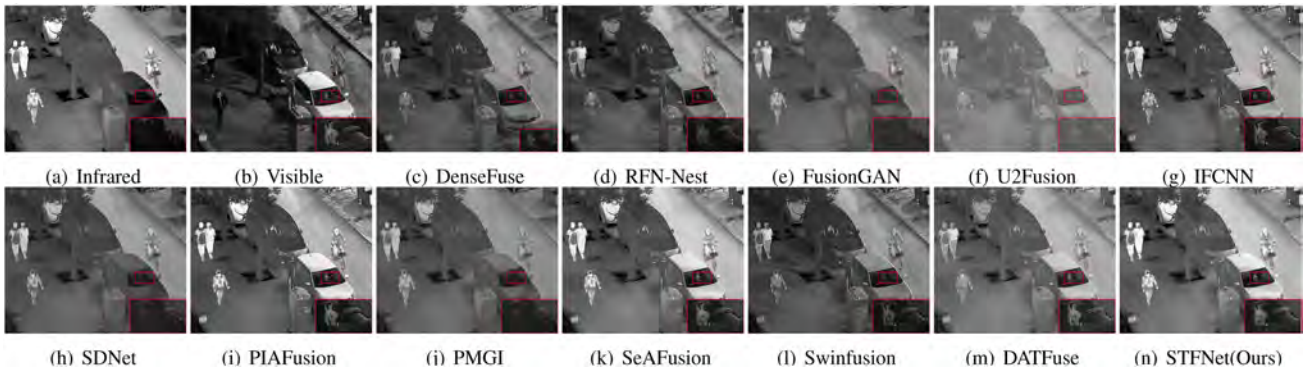

Fig. 10. Visualized comparison of the proposed fusion method (STFNet) with nine state-of-the-art methods on the '190066' image pair of the LLVIP [65] dataset.

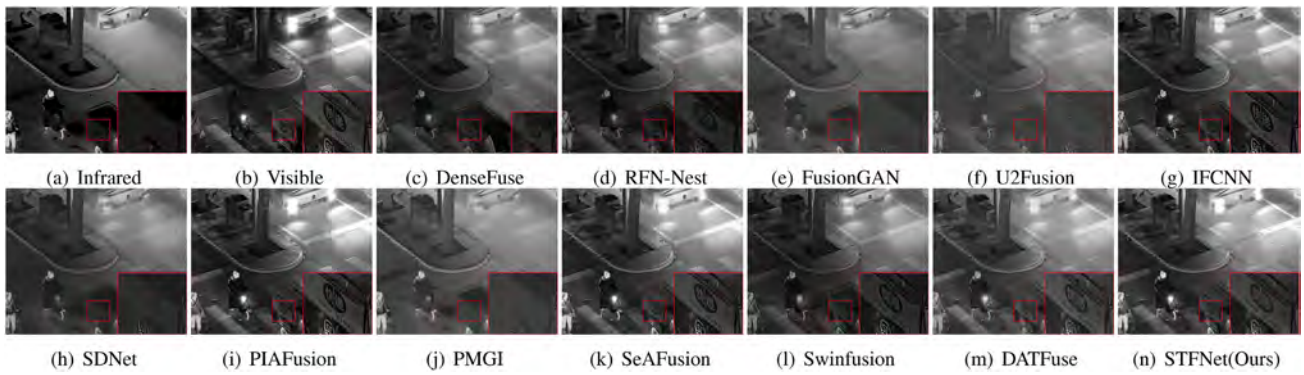

Fig. 11. Visualized comparison of the proposed fusion method (STFNet) with nine state-of-the-art methods on the '210021' image pair of the LLVIP [65] dataset.

V. Conclusion

In this paper, we present a novel one-stage self-supervised method for fusing infrared and visible images, called STFNet. Unlike existing hand-designed and convolution based fusion strategies, our method leverages the Transformer-based feature fusion network which allows us to train the encoder and decoder end-to-end for optimal fusion performance. Our proposed method utilizes multi-head self-attention and multi-head cross attention to model the contextual relationships between the infrared and visible images, which is crucial for the image fusion task. Additionally, we design a feature align network to address the slight misalignment problem that can arise during the fusion process. This module helps to effectively reduce artifacts in the fused images. To ensure end-to-end training of our fusion model, we propose a self-supervised multi-task loss that includes three loss functions: structural similarity loss, frequency consistency loss, and Fourier spectral consistency loss. These three losses work together to enhance the fusion performance from different perspectives. Extensive experiments conducted on four widely used image fusion datasets have demonstrated the effectiveness of our proposed method. The results show that STFNet achieves

Authorized licensed use limited to: University Town Library of Shenzhen. Downloaded on March 27,2024 at 04:02:13 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

competitive performance compared with existing state-of-the-art fusion methods in terms of both qualitative and quantitative evaluations. In the future, we will explore more compact and effective Transformer architecture and stronger detail constrain for getting better fused results. In addition, large-scale unaligned image fusion is more practical, and we will study a unified framework for alignment and fusion.

REFERENCES

[1] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Inf. Fusion*, vol. 45, pp. 153–178, 2019.

[2] J. Hu and S. Li, "The multiscale directional bilateral filter and its application to multisensor image fusion," *Inf. Fusion*, vol. 13, no. 3, pp. 196–206, 2012.

[3] S. Li, X. Kang, and J. Hu, "Image fusion with guided filtering," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2864–2875, Jul. 2013.

[4] H.-M. Hu, J. Wu, B. Li, Q. Guo, and J. Zheng, "An adaptive fusion algorithm for visible and infrared videos based on entropy and the cumulative distribution of gray levels," *IEEE Trans. Multimedia*, vol. 19, no. 12, pp. 2706–2719, Dec. 2017.

[5] H. Li, X.-J. Wu, and J. Kittler, "MDLatLRR: A novel decomposition method for infrared and visible image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4733–4746, 2020.

[6] G. Li, Y. Lin, and X. Qu, "An infrared and visible image fusion method based on multi-scale transformation and norm optimization," *Inf. Fusion*, vol. 71, pp. 109–129, 2021.

[7] B. Yang and S. Li, "Pixel-level image fusion with simultaneous orthogonal matching pursuit," *Inf. Fusion*, vol. 13, no. 1, pp. 10–19, 2012.

[8] H. Yin, "Sparse representation with learned multiscale dictionary for image fusion," *Neurocomputing*, vol. 148, pp. 600–610, 2015.

[9] M. Kim, D. K. Han, and H. Ko, "Joint patch clustering-based dictionary learning for multimodal image fusion," *Inf. Fusion*, vol. 27, pp. 198–214, 2016.

[10] B. Zhang, X. Lu, H. Pei, and Y. Zhao, "A fusion algorithm for infrared and visible images based on saliency analysis and non-subsampled shearlet transform," *Infrared Phys. Technol.*, vol. 73, pp. 286–297, 2015.

[11] D. P. Bavirisetti and R. Dhuli, "Two-scale image fusion of visible and infrared images using saliency detection," *Infrared Phys. Technol.*, vol. 76, pp. 52–64, 2016.

[12] Y. Yang, Y. Que, S. Huang, and P. Lin, "Multiple visual features measurement with gradient domain guided filtering for multisensor image fusion," *IEEE Trans. Instrum. Meas.*, vol. 66, no. 4, pp. 691–703, Apr. 2017.

[13] H. Jin and Y. Wang, "A fusion method for visible and infrared images based on contrast pyramid with teaching learning based optimization," *Infrared Phys. Technol.*, vol. 64, pp. 134–142, 2014.

[14] H. Jin, Q. Xi, Y. Wang, and X. Hei, "Fusion of visible and infrared images using multiobjective evolutionary algorithm based on decomposition," *Infrared Phys. Technol.*, vol. 71, pp. 151–158, 2015.

[15] J. Ma, Z. Zhou, B. Wang, and H. Zong, "Infrared and visible image fusion based on visual saliency map and weighted least square optimization," *Infrared Phys. Technol.*, vol. 82, pp. 8–17, 2017.

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

[17] H. Li and X.-J. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2614–2623, May 2019.

[18] H. Li, X.-J. Wu, and J. Kittler, "Infrared and visible image fusion using a deep learning framework," in *Proc. 24th Int. Conf. Pattern Recognit.*, 2018, pp. 2705–2710.

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

[20] H. Xu, J. Ma, Z. Le, J. Jiang, and X. Guo, "Fusiondn: A unified densely connected network for image fusion," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 07, pp. 12484–12491.

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

[22] J. Ma, H. Zhang, Z. Shao, P. Liang, and H. Xu, "GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 70, 2021, Art. no. 5005014.

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

[24] R. Hou et al., "VIF-Net: An unsupervised framework for infrared and visible image fusion," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 640–651, 2020.

[25] 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 Trans. Instrum. Meas.*, vol. 70, 2021, Art. no. 5009513.

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

[27] 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 *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12797–12804.

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

[29] H. Xu, H. Zhang, and J. Ma, "Classification saliency-based rule for visible and infrared image fusion," *IEEE Trans. Comput. Imag.*, vol. 7, pp. 824–836, 2021.

[30] 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 Trans. Instrum. Meas.*, vol. 69, no. 12, pp. 9645–9656, Dec. 2020.

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

[32] 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 Trans. Image Process.*, vol. 30, pp. 4070–4083, 2021.

[33] Z. Wang, Y. Chen, W. Shao, H. Li, and L. Zhang, "SwinFuse: A residual swin transformer fusion network for infrared and visible images," *IEEE Trans. Instrum. Meas.*, vol. 71, 2022, Art. no. 5016412.

[34] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.

[35] P. Li, "DIDFuse: Deep image decomposition for infrared and visible image fusion," in *Proc. 29th Int. Conf. Int. Joint Conf. Artif. Intell.*, 2021, pp. 976–976.

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

[37] L. Qu et al., "TransFuse: A unified transformer-based image fusion framework using self-supervised learning," 2022, *arXiv:2201.07451*.

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

[39] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 502–518, Jan. 2022.

[40] W. Tang, F. He, Y. Liu, Y. Duan, and T. Si, "DATFuse: Infrared and visible image fusion via dual attention transformer," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 7, pp. 3159–3172, Jul. 2023.

[41] J. Li, J. Zhu, C. Li, X. Chen, and B. Yang, "CGTF: Convolution-guided transformer for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 71, 2022, Art. no. 5012314.

[42] 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 J. Automatica Sinica*, vol. 9, no. 7, pp. 1200–1217, Jul. 2022.

[43] Z. Chang, Z. Feng, S. Yang, and Q. Gao, "AFT: Adaptive fusion transformer for visible and infrared images," *IEEE Trans. Image Process.*, vol. 32, pp. 2077–2092, 2023.

[44] H. Xu, J. Ma, J. Yuan, Z. Le, and W. Liu, "RFNet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 19679–19688.

[45] D. Wang, J. Liu, X. Fan, and R. Liu, "Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration," in *Proc. Int. Joint Conf. Artif. Intell.*, 2022, pp. 3508–3515.

[46] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 1–11.

Authorized licensed use limited to: University Town Library of Shenzhen. Downloaded on March 27,2024 at 04:02:13 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

[47] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–14.
[48] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.
[49] B. Yan, H. Peng, J. Fu, D. Wang, and H. Lu, "Learning spatio-temporal transformer for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 10448–10457.
[50] N. Wang, W. Zhou, J. Wang, and H. Li, "Transformer meets tracker: Exploiting temporal context for robust visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1571–1580.
[51] R. Guo, D. Niu, L. Qu, and Z. Li, "Sotr: Segmenting objects with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 7157–7166.
[52] X. Chen, B. Yan, J. Zhu, D. Wang, X. Yang, and H. Lu, "Transformer tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8126–8135.
[53] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9729–9738.
[54] I. Misra and L. v. d. Maaten, "Self-supervised learning of pretext-invariant representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6707–6717.
[55] D. Yuan, X. Chang, P.-Y. Huang, Q. Liu, and Z. He, "Self-supervised deep correlation tracking," *IEEE Trans. Image Process.*, vol. 30, pp. 976–985, 2021.
[56] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.
[57] S. Gidaris, A. Bursuc, N. Komodakis, P. Pérez, and M. Cord, "Boosting few-shot visual learning with self-supervision," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8059–8068.
[58] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2015, pp. 234–241.
[59] S.-Y. Cao, H.-L. Shen, S.-J. Chen, and C. Li, "Boosting structure consistency for multispectral and multimodal image registration," *IEEE Trans. Image Process.*, vol. 29, pp. 5147–5162, 2020.
[60] M. Arar, Y. Ginger, D. Danon, A. H. Bermano, and D. Cohen-Or, "Unsupervised multi-modal image registration via geometry preserving image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13410–13419.
[61] J. Dai et al., "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 764–773.
[62] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1037–1045.
[63] J. Liu et al., "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5802–5811.
[64] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, 2003, vol. 2, pp. 1398–1402.
[65] X. Jia, C. Zhu, M. Li, W. Tang, and W. Zhou, "LLVIP: A visible-infrared paired dataset for low-light vision," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3496–3504.
[66] A. Ellmauthaler, C. L. Pagliari, E. A. da Silva, J. N. Gois, and S. R. Neves, "A visible-light and infrared video database for performance evaluation of video/image fusion methods," *Multidimensional Syst. Signal Process.*, vol. 30, no. 1, pp. 119–143, 2019.
[67] A. Toet, "TNO image fusion dataset," May 2014. [Online]. Available: https://figshare.com/articles/dataset/TNO_Image_Fusion_Dataset/1008029
[68] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, "Image fusion meets deep learning: A survey and perspective," *Inf. Fusion*, vol. 76, pp. 323–336, 2021.
[69] H. Zhang and J. Ma, "SDNet: A versatile squeeze-and-decomposition network for real-time image fusion," *Int. J. Comput. Vis.*, vol. 129, no. 10, pp. 2761–2785, 2021.
[70] 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," *Inf. Fusion*, vol. 82, pp. 28–42, 2022.

Qiao Liu received the Ph.D. degree in computer science from the Harbin Institute of Technology (Shenzhen), Shenzhen, China, in 2021. He is currently an Associate Professor with the National Center for Applied Mathematics, Chongqing Normal University, Chongqing, China. His research interests include thermal infrared object tracking, infrared image processing, and image fusion.

Jiatian Pi received the Ph.D. degree in information and communication engineering from the Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai, China, in 2017. He is currently an Associate Professor with the National Center for Applied Mathematics, Chongqing Normal University, Chongqing, China. His research interests include computer vision, machine learning, and adaptive traffic signal control.

Peng Gao (Member, IEEE) received the Ph.D. degree from the Harbin Institute of Technology, Harbin, China, in 2020. He is currently an Associate Professor with the School of Cyber Science and Engineering, Qufu Normal University, Jining, China. His research interests include signal and image processing, deep learning, and computer vision. He is a Reviewer of top journals, such as IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, *Information Sciences*, *Knowledge-Based Systems*, *Neurocomputing*, and *IET Image Processing*.

Di Yuan (Member, IEEE) received the Ph.D. degree in computer applied technology from the Harbin Institute of Technology, Harbin, China, in 2021. He is currently a Lecturer with the Guangzhou Institute of Technology, Xidian University, Guangzhou, China. He was sponsored by China Scholarship Council as a Visiting Ph.D. Student with the Faculty of Information Technology, Monash University Clayton Campus, Melbourne, VIC, Australia, from 2019 to 2021, working with Prof. Xiaojun Chang. His research interests include object tracking, machine learning, self-supervised learning, and active learning.

Authorized licensed use limited to: University Town Library of Shenzhen. Downloaded on March 27,2024 at 04:02:13 UTC from IEEE Xplore. Restrictions apply.