Information Fusion 95 (2023) 26–41

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

Feature dynamic alignment and refinement for infrared–visible image fusion: Translation robust fusion

Huafeng Li [a,1], Junzhi Zhao [a,1], Jinxing Li [b,c,*], Zhengtao Yu [a], Guangming Lu [b,c]

[a] Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Yunnan, Kunming, 650500, PR China
[b] Harbin Institute of Technology, Shenzhen, PR China
[c] Shenzhen Key Laboratory of Visual Object Detection and Recognition, Shenzhen, PR China

ARTICLE INFO

Keywords:
Image fusion
Visible image
Infrared image
Feature-alignment
Translational misalignment

ABSTRACT

Translational displacement between source images from different sensors is a general phenomenon, which will cause performance degradation on image fusion. To tackle this issue, a straightforward way is to make source images registration first. However, due to the large modality-gap between the infrared image and the visible image, it is too challenging to achieve completely registered images. In this paper, a novel registration-free fusion method is primarily proposed for infrared and visible images with translational displacement, which transforms the problem of image registration to feature alignment in an end-to-end framework. Specifically, we propose a cross-modulation strategy followed by feature dynamic alignment, so that the spatial correlation of shifts is adaptively measured and the aligned features can be dynamically extracted. A feature refinement module is additionally designed based on the local similarity, which enhances the textures related information while suppresses artifacts related information. Thanks to these strategies, our experimental results on infrared–visible images with translational displacement achieve dramatic enhancement compared with state-of-the-arts. To the best of our knowledge, this is the first work on infrared–visible image fusion without strict registration. It does break the constraint of existing image-registration based two-step strategies and provide a simple but efficient way for multi-modal image fusion. The source code will be released at https://github.com/lhf12278/RFVIF.

1. Introduction

By capturing the reflected light, a visible image can be generated, which enjoys abundant textural and structural information but is sensitive to the illumination. By contrast, the infrared image captures the thermal radiation, being robust to the light but lack of details. Due to the complementary information between these two types of images, the infrared–visible image fusion is studied, through which a fused image is obtained that not only enjoys detailed textures but also contains salient objects. Thanks to these merits, infrared–visible image fusion is widely applied to object detection, recognition, and tracking in surveillance [1], vehicle navigation [2] and monitoring [3].

By the grace of the development of deep learning, a number of deep learning based infrared–visible image fusion methods have been proposed in recent years [4–13], which do gain technological advancement. In particular, Liu et al. [9] first proposed a pyramid network, effectively fusing the visible image and the infrared image through an end-to-end strategy. In addition, Li et al. [8] proposed the fusion framework of infrared and visible images with different resolutions, realizing the fusion of source images from different resolutions to arbitrary resolutions. Xu et al. [11–13] unified multi-source image fusion into one framework. Ma et al. [14] introduced the saliency target detection into the fusion of infrared and visible images, which effectively preserved the information of the targets in the infrared image.

However, the aforementioned methods are limited to a strong constraint that the pair of images is completely registered pixel-by-pixel. Since visible and infrared images are captured by different sensors, the spatial disparity between two sensors is inevitable, which subsequently results in translational misalignment between source images. Besides, the different electronic magnification of two sensors will also commonly cause translational misalignment. If existing fusion approaches are directly applied to misaligned images, there would be a low-quality or even valueless fused image. To tackle this issue, a naive way is to make registration first and these preprocessed images are then fused by

* Corresponding author at: Harbin Institute of Technology, Shenzhen, PR China.
E-mail address: lijinxing158@gmail.com (J. Li).
[1] Equal contribution.

https://doi.org/10.1016/j.inffus.2023.02.011
Received 2 January 2023; Received in revised form 4 February 2023; Accepted 7 February 2023
Available online 10 February 2023

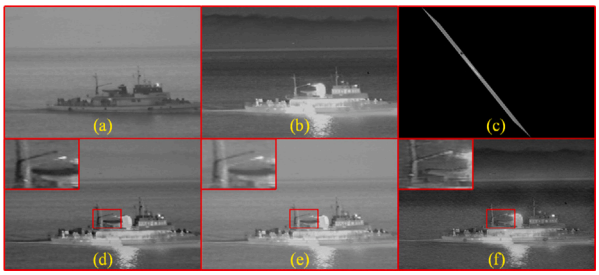

Fig. 1. Fusion results of different methods on images with translational misalignment. (a) and (b) are source images with shifts of 5 pixels, (c) is the registered infrared image using HAPCG [22], (d) is the fusion result of NestFuse [23], (e) is the fusion result of SwinFusion [24], (f) is the fusion result of our method.

exploiting fusion algorithms [15–19], or some researchers pay attention to combine the image registration and the image fusion into a united model [10,20,21]. It is true that these two-step techniques seem to be a reasonable solution, while it may cause the error delivery, which means the quality of fusion depends on the accuracy of the registration. As displayed in Fig. 1(c), the state-of-the-art infrared–visible image registration method HAPCG [22] is introduced to make registration for the source images with translational misalignment in Fig. 1(a) and Fig. 1(b). However, it is easy to observe that the infrared image is collapsed to a line. In other words, the infrared image does not make any contribution to fusion. By contrast, if the images are fused directly without registration, the boundaries of objects meet heavy artifacts, as visualized from Fig. 1(d) to (e). The explicit registration process may be rational and essential when the form of misalignment is complex (e.g., rotation and elastic transformation). However, it is definitively not a cost-efficient way on fusing these source images with translational misalignment due to the risk of registration failure. Thus, to further apply infrared–visible image fusion to practical applications, it is really significant to study a novel method which is capable of avoiding pixel-level based strict registration but fusing two images simultaneously.

In this paper, we propose a novel shifts robust fusion method for infrared and visible images. To the best of our knowledge, this is the first one which is completely free from strict registration of images, gaining the breakthrough in multi-modal image fusion. In our work, we transform the pixel-level based image registration to feature alignment during feature extraction, so that the aligned features can be then fused and reconstructed to a high-quality image. Thanks to this technique, we do not need to register the source images strictly. Specifically, we take the visible features as references, and then design a cross modulation mechanism based on the awareness of the offset between two source images. By perceiving their latent relationship in the spatial space, this mechanism generates modulation parameters to modulate features. Followed by the feature dynamic alignment module, infrared features are encouraged to be spatially consistent to the visible features according to the adaptively learned convolutional kernels.

In the feature reconstruction phase, a multi-grained feature refinement module is further proposed to enhance the guidance of the reference image and suppress the influence of artifacts on our fused image according to the local-similarity. In detail, features which are similar to the visible modality are further enhanced, so that more textural information is preserved. Meanwhile, infrared features that are not well aligned in the early processing are suppressed, so that the artifacts caused by the misalignment between source images are alleviated. Fig. 1(f) displays the image fused by our method when the visible image and the infrared image encounter five shifting pixels on both horizontal and vertical axis. It is easy to see that our presented method effectively alleviates the problem on mismatching and removes artifacts in the fused image in comparison to other existing state-of-the-arts.

The contributions of this paper are concluded as follows:

- To the best of our knowledge, we are the first one who propose a novel fusion method for infrared and visible images with translational misalignment, which is completely free from strict registration. This strategy is capable of removing the artifacts in the fused image without accurate registration, which can avoid the influence of registration failure on the fusion results. We believe this work will give a valuable inspiration for other researchers on the study of multi-modal image fusion.
- A cross-modulation mechanism followed by a feature dynamic alignment module is presented. These two techniques reveal the latent relationship between two source images in the spatial space and subsequently generate modulation parameters and dynamic convolutional kernels, which adaptively align the infrared features to the visible features, preventing from the image registration. Additionally, a local-similarity guided feature refinement strategy is studied to further enhance the guidance of the reference image and suppress the influence of artifacts on reconstructed fused image.
- Experimental results on benchmarks substantiate the dramatic effectiveness and superiority of our method compared with both hand-crafted and deep learning based fusion methods. Particularly, images fused by our method not only removes the artifacts, but also gains abundant textural details.

The rest of this paper is organized as follows. In Section 2, some related works about image fusion and image registration are briefly described. Our proposed registration-free method is then analyzed in Section 3. Experiments are conducted in Section 4 to show our superiority quantitatively and qualitatively, followed by the conclusion in Section 5.

2. Related works

2.1. Infrared–visible image fusion

Followed by the development of deep learning, many deep learning based image fusion methods have been studied [25]. In 2017, Liu et al. [26] first proposed a CNN based multi-focus image fusion method, which decomposed an image into multiple patches and classified them to the focused or defocused class, preventing from designing hand-crafted features. Inspired by it, deep learning is then applied to infrared–visible image fusion. Liu et al. [9] proposed a pyramid-like network, which exploited both textural and semantic information from multi-scale features. Similarly, Li et al. [27] developed DenseFuse network to densely exploit features from multiple layers, achieving more detailed information in the fused image. Inspired by image decomposition, Lahoud et al. [28] presented real-time image fusion method. In this method, the two source images are separated into different components followed by different fusion rules, so that more interested information was preserved.

Besides, a bilevel integrated model was proposed by Liu et al. [29], which achieved fusion via an alternative optimization. Jian et al. [30] designed a symmetric encoder–decoder for infrared and visible image fusion. Due to the lack of the labeled data, Jung et al. [31] proposed an unsupervised loss function using the structure tensor representation of the multi-channel image contrasts. Also, end-to-end encoder–decoder networks RFN-Nest [32] and NestFuse [23] were respectively studied, avoiding designing the fusion rule based on human's prior. Thanks for the powerful capability of image generation, GAN [33] has been widely applied to infrared–visible image fusion. For instance, Ma et al. [34] first exploited GAN to generate the image from two source inputs, while this method meet a limitation on textural detail preservation. Thus, Ma et al. [35,36] made improvement on [34] to further enhance the quality of fused images. And some high-level tasks driven fusion methods [37–40] emerged to promote the segmentation and object

detection tasks. Additionally, some more general models [11,13,24,41–43] were proposed which were adaptive for multiple types of image fusion tasks.

Although aforementioned approaches gain satisfactory performance, they all follow the constraint that two source images are completely registered pixel-by-pixel. If there is translational misalignment between source images, existing methods will meet the artifacts in the fused result.

2.2. Multi-modal image registration

Due to different spatial locations and resolutions of different sensors, it is inevitable that images from multiple modalities are misaligned [44,45]. Therefore, some researchers have tried to tackle this issue. Lv et al. [46] jointly took the gradient feature and the SIFT feature into account for image registration. To further enhance the accuracy and reduce the algorithm complexity, the speed up robust feature based on SIFT was also learned in [47,48]. By introducing PCA, the dimension of the SIFT feature is decreased, subsequently improving the efficiency on speed and restoration [49]. To increase the number of correct correspondences, Ma et al. [50] proposed an enhanced feature matching approach, which jointly considered the position, scale, and orientation of each keypoint. To be adaptive for video tasks, Rosten et al. [51] studied a novel corner detection method named feature from accelerated segment test (FAST), remarkably accelerating the speed. By combining the orientation, FAST was extended to ORB [52], which further increased the robustness. To remove mismatches, Ma et al. [53] developed a method termed as locality preserving matching by persisting the local structures of those features matched correctly. Besides, inspired by the consistence of phases, some related strategies were also learned for multi-modal image fusion, e.g., consistent local phase representation [54], radio invariant transformation [55], and phase correlation based on Log-Gabor filtering [56]. To further improve the accuracy, some researches focused on removing the mismatched features to establish reliable feature correspondences [57–59].

Despite the fact that image registration methods described above do meet the requirement for some tasks, they have the probability of failing to pair the infrared images and visible images for the sake of the large modality-gap and the low-quality of source images. Furthermore, in the matching phase, it is a general phenomenon that there is distortion on textures and structures etc. [60], which decreases the quality of the fused image. Thus, the registration process is not the first choice for the slightly misaligned images due to these limitations mentioned above. Therefore, a shifts robust infrared–visible image fusion method is indeed significant.

3. The proposed method

3.1. Overview

The pipeline of our proposed method is shown in Fig. 2, which consists of cross-modulation feature extraction module (CMFEM), feature dynamic alignment module (FDAM), multi-grained feature refinement module (MGFRM), and pyramid feature fusion module (PFFM). In CMFEM, the cross-modulation strategy is embedded which aims to extract the latent relationship between two source image features, generating modulation parameters to modulate these two features. According to the modulated features, the dynamic convolutional kernels are generated for the infrared image via FDAM, which are then used to align the infrared features to the visible features spatially. Followed by MGFRM, the obtained features under multiple scales are further refined by enhancing the textural information in the visible modality but suppressing inconfidential information in the infrared modality based on a similarity metric. Finally, we fuse multiple hierarchical features and obtain the fused image through PFFM and residual blocks respectively, which is free from artifacts and enjoys abundant textures.

3.2. Cross-modulation feature extraction module

As analyzed in [61], features in lower layers enjoy spatial details, while features in deeper layers gain more semantic information. To exploit their complementary information, here we design the multi-scale residual block (MSRB) based pyramid feature extractor. As shown in Fig. 2(b), the MSRB jointly combines the features with different receptive fields, so that richer features are extracted. From Fig. 2 (left), we can see that our module consists of three parts. The first part is composed of three MSRBs followed by a cross-modulation strategy, and the second and the third parts both contain two MSRBs and a cross-modulation strategy. According to this module, the features of two source images can be obtained. Mathematically, denote the misaligned visible and infrared images to be $\mathbf{x}_{vi} \in \mathbb{R}^{H\times W}$ and $\mathbf{x}_{ir} \in \mathbb{R}^{H\times W}$ respectively, where H and W are the hight and width of the source image. By forwarding them to three MSRBs in the first part, we can get

$$\mathbf{F}_{1,vi} = D_1(\mathbf{x}_{vi}),\ \mathbf{F}_{1,ir} = D_2(\mathbf{x}_{ir}), \tag{1}$$

where D_1 and D_2 are both composed of the three MSRBs in the first part. Similarly, ($\mathbf{F}_{2,vi}$, $\mathbf{F}_{2,ir}$) and ($\mathbf{F}_{3,vi}$, $\mathbf{F}_{3,ir}$) can also be computed from the second and third parts.

Instead of making registration for the two source images pixel-by-pixel, in this paper we focus on aligning their high-level features, so that the aligned feature can be fused and reconstructed to an image without artifacts. Since any pairs of images meet the different shifting pixels on position and orientation, an encoder with fixed parameters is incapable of modeling these various cases. To address this issue, a reasonable way is to dynamically extract features according to their latent relationship in the spatial space. Thus, we introduce a cross-modulation strategy, which is professional in revealing the latent relationship and then generating a set of dynamic modulation parameters. Based on dynamic parameters, features, e.g., $\mathbf{F}_{1,vi}$ and $\mathbf{F}_{1,ir}$, are cross-modulated, being greatly beneficial for perfect feature alignment in following steps.

The architecture of cross-modulation is shown in Fig. 3, which contains convolutional layers and activation functions. Denote $\mathbf{F}_{l,vi}$ and $\mathbf{F}_{l,ir}$ ($l = \{1, 2, 3\}$) as the outputs from MSRBs of the aforementioned three parts. By forwarding them into a convolutional layer, their outputs are concatenated.

$$\mathbf{F}^{c}_{l,vi+ir} = \mathrm{concat}(\mathrm{conv}(\mathbf{F}_{l,vi}, 3\times 3, 2), \mathrm{conv}(\mathbf{F}_{l,ir}, 3\times 3, 2)), \tag{2}$$

where 3×3 is the size of the convolutional kernel and 2 is the stride. In order to gain the transformation for modulating the visible and infrared features, we then pass $\mathbf{F}^{c}_{l,vi+ir}$ through three convolutional layers and averagely split the obtained feature maps according to channels:

$$(\mathbf{F}^{vi}_{l,vi+ir}, \mathbf{F}^{ir}_{l,vi+ir}) = \mathrm{split}(\mathrm{convs}(\mathbf{F}^{c}_{l,vi+ir}, 3\times 3)). \tag{3}$$

Based on Eq. (4), the modulation parameters for two modalities are computed.

$$(\boldsymbol{\gamma}_{l,i}, \boldsymbol{\beta}_{l,i}) = \mathrm{sigmoid}(\mathrm{convs}(\mathbf{F}^{i}_{l,vi+ir}, 3\times 3), \mathrm{convs}(\mathbf{F}^{i}_{l,vi+ir}, 3\times 3)),\quad s.t.,\ i = (vi, ir) \tag{4}$$

Finally, we can get the modulated feature for each modality as follows:

$$\mathbf{F}^{cm}_{l,i} = \mathrm{conv}(\mathbf{F}_{l,i}, 3\times 3, 2) \odot \boldsymbol{\gamma}_{l,i} + \boldsymbol{\beta}_{l,i},\ s.t.,\ i = (vi, ir), \tag{5}$$

where $\odot$ means the hadamard product. In fact, Eq. (5) involves the theory of affine transformation, which has been widely used in image processing [62–64]. Thanks to Eq. (5), the visible and infrared features are both transformed in scale and shifting numerically, subsequently contributing to feature alignment in the following feature dynamic alignment module.

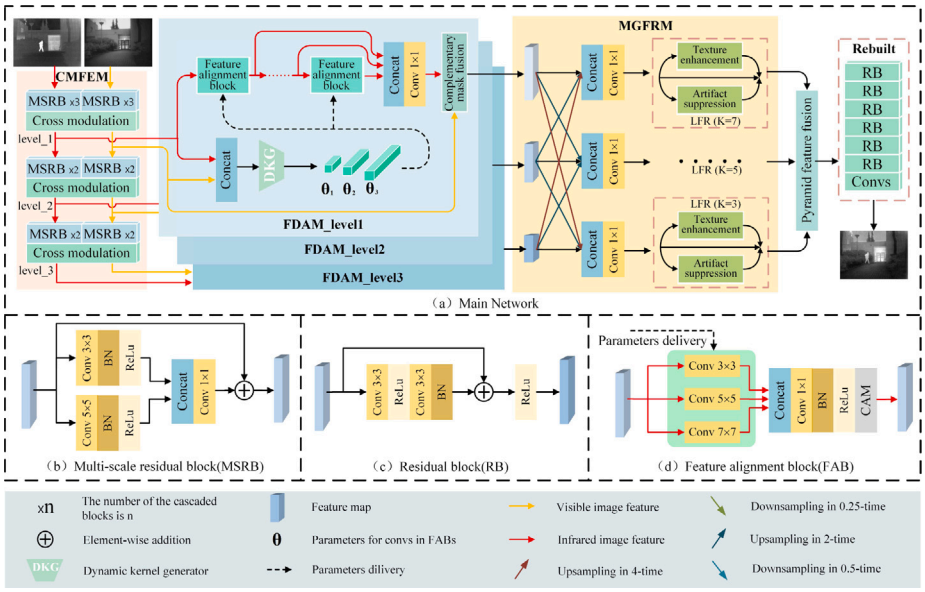

Fig. 2. Overview of the proposed architecture. It consists of cross-modulation feature extraction module (CMFEM), feature dynamic alignment module (FDAM), multi-grained feature refinement module (MGFRM), and pyramid feature fusion module (PFFM). Thanks to CMFEM and FDAM, the cross-modulation parameters and dynamic kernels are generated according to the inputting pair, so that the infrared features are enforced to be aligned with the visible features. To further refine the obtained features both in alignment and textural information enhancement, MGFRM is followed. Finally, PFFM is exploited to fused multiple features from various layers to reconstruct the fused image.

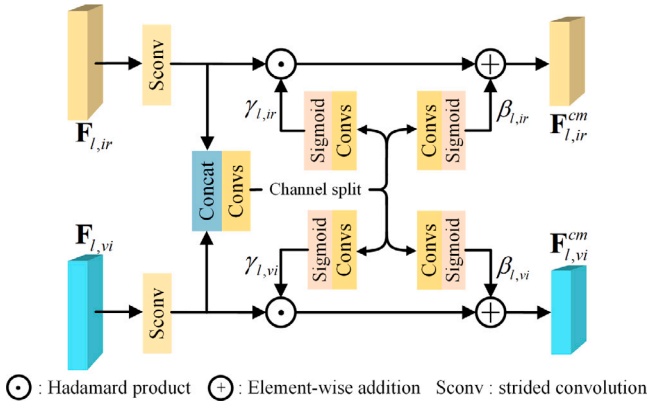

Fig. 3. The architecture of the cross-modulation strategy.

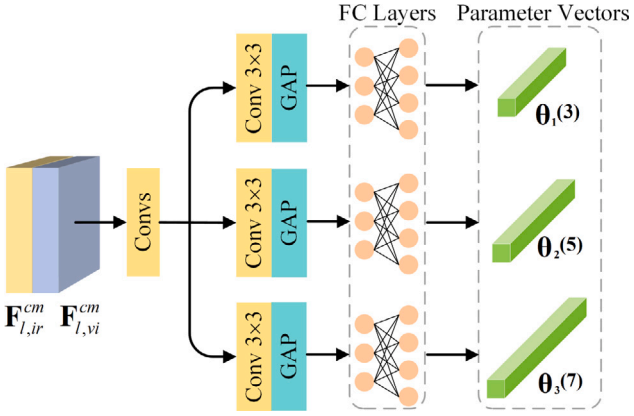

Fig. 4. The architecture of the dynamic kernel generator.

3.3. Feature dynamic alignment module

The feature maps are obtained by summarizing elements in their receptive fields via a weighted way. Thus, the convolutional kernel plays a significant role for feature alignment. Existing works encounter fixed kernels if the model has been trained. As mentioned above, different pairs encounter diverse cases of translational misalignment. Of course, convolutional kernels with fixed parameters cannot meet our requirement. Instead of learning kernels with fixed parameters, we propose a shifting perception based dynamic kernel learning strategy, so that parameters in the kernel are estimated according to different pairs of modulated features automatically.

In detail, the dynamic kernel generator (DKG) is shown in Fig. 4. It is easy to observe that three dynamic kernels under different scales are generated from three branches, being robust for diverse cases of translational misalignment. Mathematically, by forwarding the concatenation of $\mathbf{F}_{l,vi}^{cm}$ and $\mathbf{F}_{l,ir}^{cm}$ through each branch B_k (a convolutional layer, global average pooling operator and a fully connected layer are contained in each B_k), we can get its associated dynamic kernel

$$\theta_k(2k+1) = B_k(\text{convs}(\text{concat}(\mathbf{F}_{l,vi}^{cm}, \mathbf{F}_{l,ir}^{cm}))), \tag{6}$$

where $\theta_k(2k+1)$ denotes the size of the dynamic kernel is $(2k+1)\times(2k+1)$. From Eq. (6), we can observe that $\theta_k(2k+1)$ is computed based on our two types of modulated features, so that it is capable of modeling their misalignment.

By making convolutional for infrared features according to our learned dynamic kernels, the infrared features can be well aligned to visible features. Particularly, assume the kernel as conv_{θ_k}, whose parameters are $\theta_k(2k+1)$. As shown in Fig. 2(d), the **feature alignment block (FAB)** is constructed by combining conv_{θ_k}, concat, conv(1×1), BN, activation function ReLu and channel attention module (CAM). By concatenating feature maps obtained by three dynamic convolution kernels, conv(1×1) is exploited to fuse the multi-scale features, and the fused feature is then normalized for CAM, enhancing the importance of different channels. Furthermore, as displayed in Fig. 2(a), we adopt multiple feature alignment blocks one by one, so that the infrared features are gradually refined to gain better alignment. Finally, taking the visible modality as the reference, we then get the aligned infrared feature $\mathbf{F}_{l,ir}^{align}$ and its corresponding reconstructed image $\mathbf{x}_{l,ir}^{align}$ by utilizing the conv(1×1) mapping whose input is the combination of multiple aligned features from different feature alignment blocks.

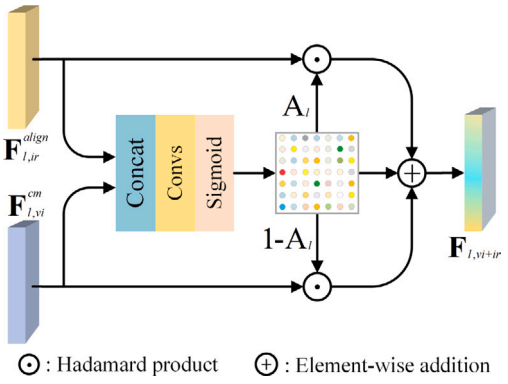

Fig. 5. The architecture of the complementary mask fusion strategy.

Of course, we prefer $\mathbf{x}_{l,ir}^{align}$ to completely pair with $\hat{\mathbf{x}}_{l,ir}$, where $\hat{\mathbf{x}}_{l,ir}$ is the ground-truth which is already aligned with $\mathbf{x}_{vi}$ under different image sizes. To achieve this issue, the l_1-loss is introduced:

$$L_{align} = \sum_{l=1}^{3} \left\| \hat{\mathbf{x}}_{l,ir} - \mathbf{x}_{l,ir}^{align} \right\|_1. \tag{7}$$

After getting $\mathbf{F}_{l,vi}^{cm}$ and its aligned infrared feature $\mathbf{F}_{l,ir}^{align}$, feature fusion should be followed to gain the feature which contains both textual and thermal information. A naive way is to concatenate or add two types of features. Although this strategy is easy to be implemented, pixels that we are interested in are not well measured, limiting the quality of the fused image. Besides, the redundant mutual information in the source images will also confuse the learning of networks. Thus, here we propose a complementary mask fusion (CMF) strategy, which allows us to transform our interested information from two modalities into the fused image. Mathematically, as displayed in Fig. 5, we first concatenate $\mathbf{F}_{l,vi}^{cm}$ and $\mathbf{F}_{l,ir}^{align}$, and then input them through three 3×3 convolutional layers and a sigmoid activation function, so that an attention map $\mathbf{A}_l$ is subsequently obtained. According to Eq. (8), fused features under different scales are achieved:

$$\mathbf{F}_{l,vi+ir} = \mathbf{F}_{l,ir}^{align} \odot \mathbf{A}_l + \mathbf{F}_{l,vi}^{cm} \odot (1 - \mathbf{A}_l). \tag{8}$$

3.4. Multi-grained feature refinement module

Following CMFEM and FDAM, we further introduce a multi-grained feature refinement module (MGFRM), which makes the similarity measurement between the features from source images and fused features in a local area, so that artifacts from the infrared image are further removed and the textural information from the visible image is also enhanced. From Fig. 2(a) we can see that, to further combine the details and semantic information of the fused features in different levels, we first fuse the features $\{\mathbf{F}_{1,vi+ir}, \mathbf{F}_{2,vi+ir}, \mathbf{F}_{3,vi+ir}\}$ to get $\mathbf{F}_{l,vi+ir}^{f} \in \{\mathbf{F}_{1,vi+ir}^{f}, \mathbf{F}_{2,vi+ir}^{f}, \mathbf{F}_{3,vi+ir}^{f}\}$, so that $\mathbf{F}_{l,vi+ir}^{f}$ can be regarded as the inputting feature map of our local feature refinement (LFR) strategy. As shown in Fig. 6, LFR consists of two branches associated with visible and infrared modalities, respectively. In our method, the visible image is regarded as the reference (benchmark) which not only guides the alignment of infrared features but also contributes textural information for the fused features. For the visible modality, the fused feature is reconstructed based on region similarity between the visible features and fused features to enhance the textural related features. For the infrared modality, since the edges of the original infrared images are highly related to the artifacts, we utilize the gradient map of the infrared image to suppress artifacts in the fused image by measuring the similarity between its gradient features and fused features.

Adaptive Visible Feature Refinement. For the adaptive visible feature refinement branch, we measure the local similarity between visible features and fused features for better reconstruction. Denote the features acquired by three 3×3 convolutional layers from visible image and its gradient map as $\mathbf{F}_{l,vi}^{e}$. And the feature vectors at each spatial position (i, j) of $\mathbf{F}_{l,vi}^{e}$ is denoted as $\mathbf{f}_{l,vi}^{ij}$. To measure the local similarity, we find the corresponding $k \times k$ patch centered at the same position (i, j) from the fused feature $\mathbf{F}_{l,vi+ir}^{f}$. For each feature vector $\mathbf{f}_{l,vi+ir}^{ij}(i+x, j+y)$ (where $x, y \in \{\frac{-k+1}{2}, \ldots, 0, \ldots, \frac{k-1}{2}\}$) in this patch, we calculate the similarity with respect to $\mathbf{f}_{l,vi}^{ij}$ as:

$$\mathbf{s}_l^{ij}(x, y) = \frac{(\mathbf{f}_{l,vi+ir}^{ij}(i+x, j+y))^T \mathbf{f}_{l,vi}^{ij}}{\left\| \mathbf{f}_{l,vi+ir}^{ij}(i+x, j+y) \right\|_2 \left\| \mathbf{f}_{l,vi}^{ij} \right\|_2}. \tag{9}$$

After $k \times k$ times measurement, we can get the similarity map $\mathbf{s}_l^{ij}$ at each position (i, j), which is the size of $k \times k$. Due to the abundant texture in the visible modality, the larger similarity value means the more textural information does $\mathbf{f}_{l,vi+ir}^{ij}(i+x, j+y)$ have. To re-weight the proffer of the $k \times k$ feature vectors in the patch when reconstructing the new feature vector at position (i, j), a Softmax function is used to generate the weights of each vectors in the patch as:

$$\mathbf{w}_l^{ij} = \text{softmax}\left(\mathbf{s}_l^{ij}\right). \tag{10}$$

Based on the weight map, the feature vectors at each position (i, j) can be reconstructed as:

$$\tilde{\mathbf{f}}_{l,vi+ir}^{ij} = \sum_{x,y} w_l^{ij}(x, y) \mathbf{f}_{l,vi+ir}^{ij}(i+x, j+y), \tag{11}$$

which means the reconstructed feature vector at each position (i, j) is acquired from the features vectors in the $k \times k$ patch whose centroid is (i, j) in a weighted way. And the more textured feature vector in the patch contributes more to the reconstruction.

Adaptive Infrared Feature Refinement. Referring to the infrared modality, denote the features from the gradient map of the infrared image as $\mathbf{F}_{l,ir}^{e}$. Then we can also get feature vector $\mathbf{f}_{l,ir}^{ij}$ from $\mathbf{F}_{l,ir}^{e}$ in position (i, j) and feature vector $\mathbf{f}_{l,vi+ir}^{ij}$ from $\mathbf{F}_{l,vi+ir}^{f}$ in position (i, j) for similarity measurement. Different from the visible modality, as we enforce the infrared feature to spatially align to the visible feature, the larger similarity between $\mathbf{f}_{l,vi+ir}^{ij}$ and $\mathbf{f}_{e,ir}^{ij}$ inversely denotes that $\mathbf{f}_{l,vi+ir}^{ij}$ is highly associated with the artifact. Thus, as shown in Eq. (12), we use 1 to subtract the similarity to be the weight of $\mathbf{f}_{l,vi+ir}^{ij}$:

$$\tilde{\mathbf{f}}_{l,vi+ir}^{ij} = \left(1 - \frac{(\mathbf{f}_{l,ir}^{ij})^T \mathbf{f}_{l,vi+ir}^{ij}}{\left\| \mathbf{f}_{l,ir}^{ij} \right\|_2 \left\| \mathbf{f}_{l,vi+ir}^{ij} \right\|_2}\right) \mathbf{f}_{l,vi+ir}^{ij} + \mathbf{f}_{l,vi+ir}^{ij}. \tag{12}$$

Note that in Eq. (12), after getting the weighted summarization, we further add $\mathbf{f}_{l,vi+ir}^{ij}$ to avoid information loss in the subtraction.

Features maps composed of all $\tilde{\mathbf{f}}_{l,vi+ir}^{ij}$ and $\tilde{\mathbf{f}}_{l,vi+ir}^{ij}$ $(i = \{1, \ldots, H_l\}, j = \{1, \ldots, W_l\})$ are denoted as $\tilde{\mathbf{F}}_{l,vi+ir} \in \mathbb{R}^{C \times H_l \times W_l}$ and $\tilde{\mathbf{F}}_{l,vi+ir} \in \mathbb{R}^{C \times H_l \times W_l}$, respectively, which are then concatenated with $\mathbf{F}_{l,vi+ir}^{f}$. Here H_l and W_l are the height and width of feature maps at the lth level. By forwarding them into a conv(1×1), we can get the fused feature $\hat{\mathbf{F}}_{l,vi+ir}$.

3.5. Pyramid feature fusion module

To avoid missing the information caused by large-scale down-sampling and introducing redundant information caused by large-scale up-sampling, here we finally introduce a pyramid feature fusion module (PFFM) to take multi-scale features into account. As shown in Fig. 7, $\hat{\mathbf{F}}_{2,vi+ir}$ is respectively fused with $\hat{\mathbf{F}}_{1,vi+ir}$ and $\hat{\mathbf{F}}_{3,vi+ir}$, and their fused features are further fused to achieve the final fusion $\mathbf{F}_{vi+ir}$. Mathematically, $\hat{\mathbf{F}}_{1,vi+ir}$ and $\hat{\mathbf{F}}_{2,vi+ir}$ can be fused via

$$\begin{aligned} \mathbf{M}_{1,2} &= \text{sigmoid}(\text{convs}(\text{concat}(\mathcal{N}(\hat{\mathbf{F}}_{2,vi+ir}), \hat{\mathbf{F}}_{1,vi+ir}))), \\ \mathbf{F}_{vi+ir}^{1,2} &= \hat{\mathbf{F}}_{1,vi+ir} \odot \mathbf{M}_{1,2} + \mathcal{N}(\hat{\mathbf{F}}_{2,vi+ir}) \odot (1 - \mathbf{M}_{1,2}), \end{aligned} \tag{13}$$

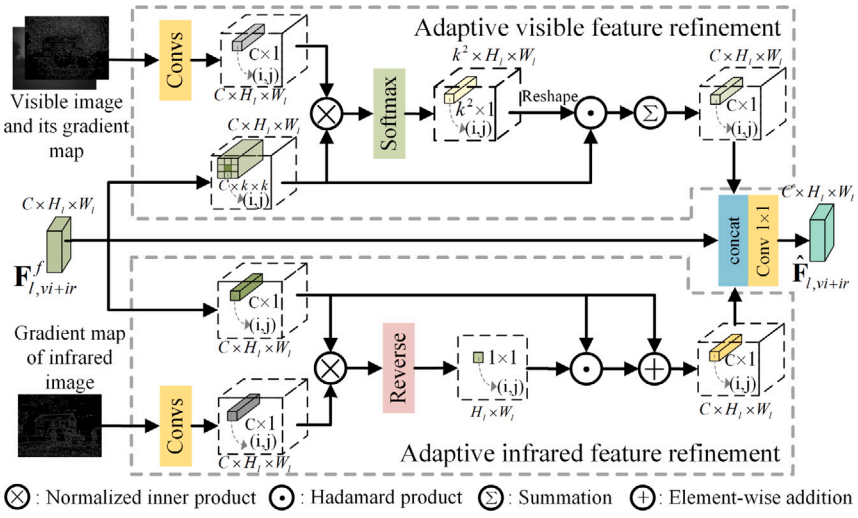

Fig. 6. The architecture of the local feature refinement strategy.

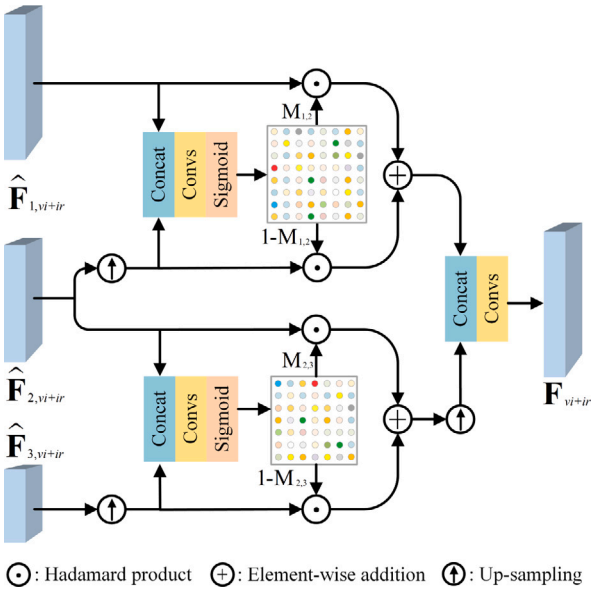

Fig. 7. The architecture of the pyramid feature fusion module.

where $\mathcal{N}(\cdot)$ means the nearest up-sampling operator. Similarly, $\hat{\mathbf{F}}_{2,vi+ir}$ and $\hat{\mathbf{F}}_{3,vi+ir}$ can be fused via

$$\begin{aligned}\mathbf{M}_{2,3} &= \text{sigmoid}(\text{convs}(\text{concat}(\mathcal{N}(\hat{\mathbf{F}}_{3,vi+ir}), \hat{\mathbf{F}}_{2,vi+ir}))),\\ \mathbf{F}^{2,3}_{vi+ir} &= \hat{\mathbf{F}}_{2,vi+ir} \odot \mathbf{M}_{2,3} + \mathcal{N}(\hat{\mathbf{F}}_{3,vi+ir}) \odot (1 - \mathbf{M}_{2,3}).\end{aligned} \tag{14}$$

Finally, $\mathbf{F}_{vi+ir}$ is obtained via Eq. (15).

$$\mathbf{F}_{vi+ir} = \text{convs}(\text{concat}(\mathcal{N}(\mathbf{F}^{2,3}_{vi+ir}), \mathbf{F}^{1,2}_{vi+ir}), 3 \times 3) \tag{15}$$

We further pass $\mathbf{F}_{vi+ir}$ through 6 residual blocks, a nearest up-sampling operator and three 3×3 convolutional layers to get the reconstructed image $\mathbf{x}_{vi+ir}$.

3.6. Loss function

Except of the alignment loss in Eq. (7), the pixel loss as well as the perception loss are also utilized.

Pixel Loss: The fused image $\mathbf{x}_{vi+ir}$ is a combination of two source images $\mathbf{x}_{vi}$ and $\mathbf{x}_{ir}$, while there is no ground-truth reference corresponding to $\mathbf{x}_{vi+ir}$. Being similar to existing methods, we exploit Eq. (16) to measure our generated image at the pixel-level.

$$l_{pixel} = (1-\alpha)\left\|\mathbf{x}_{vi+ir} - \mathbf{x}_{vi}\right\|_1 + \alpha\left\|\mathbf{x}_{vi+ir} - \hat{\mathbf{x}}_{ir}\right\|_1, \tag{16}$$

where $\hat{\mathbf{x}}_{ir}$ is the infrared image strictly registered with $\mathbf{x}_{vi}$.

Perception Loss: To encourage $\mathbf{x}_{vi+ir}$ to get salient information from $\mathbf{x}_{vi}$ and $\mathbf{x}_{ir}$, the perception loss is used as another measurement [65].

$$\begin{aligned}l_{per} = (1-\beta)&\left\|VGG16(\mathbf{x}_{vi+ir}) - VGG16(\mathbf{x}_{vi})\right\|_F\\ &+\beta\left\|VGG16(\mathbf{x}_{vi+ir}) - VGG16(\hat{\mathbf{x}}_{ir})\right\|_F,\end{aligned} \tag{17}$$

where VGG16 denotes the VGG16 network [66] pretrained on the ImageNet [67].

Finally, the total loss function of our proposed method is

$$l_{tol} = l_{pixel} + l_{per} + \lambda l_{align}, \tag{18}$$

where α, β and λ in Eq. (16), Eq. (17), and Eq. (18) are the predefined non-negative parameters.

4. Experiments

In this section, experiments are conducted to demonstrate the superiority of our proposed method. The datasets as well as implementation details are first introduced. We then describe the used metrics for quantitative evaluation. Experimental comparison between our proposed method and state-of-the-arts are then analyzed, followed by the ablation study.

4.1. Dataset construction

KAIST[2] and FLIR[3] are two widely used training datasets for infrared–visible image fusion, where KAIST consists of 95,000 pairs and FLIR contains 14,452 pairs. Here we randomly select 6,200 pairs from them as the training set, respectively, so that 12,400 pairs of images are totally used. Since these pairs have been completely registered pixel-by-pixel, which does not follow our assumption, we crop a patch with the 128×128 size from the infrared or visible images under different positions to simulate the translational misalignment. These two patches can be regarded as the training pair. Additionally, we re-crop a patch from the infrared image whose position is the same to that of the visible image. This re-cropped patch followed by the visible patch are exploited as the registered ground-truths. Furthermore, being similar to

[2] https://soonminhwang.github.io/rgbt-ped-detection/
[3] https://www.flir.ca/oem/adas/adas-dataset-form/

Fig. 8. 16 pair test images from TNO and VOT2020-RGBT datasets.

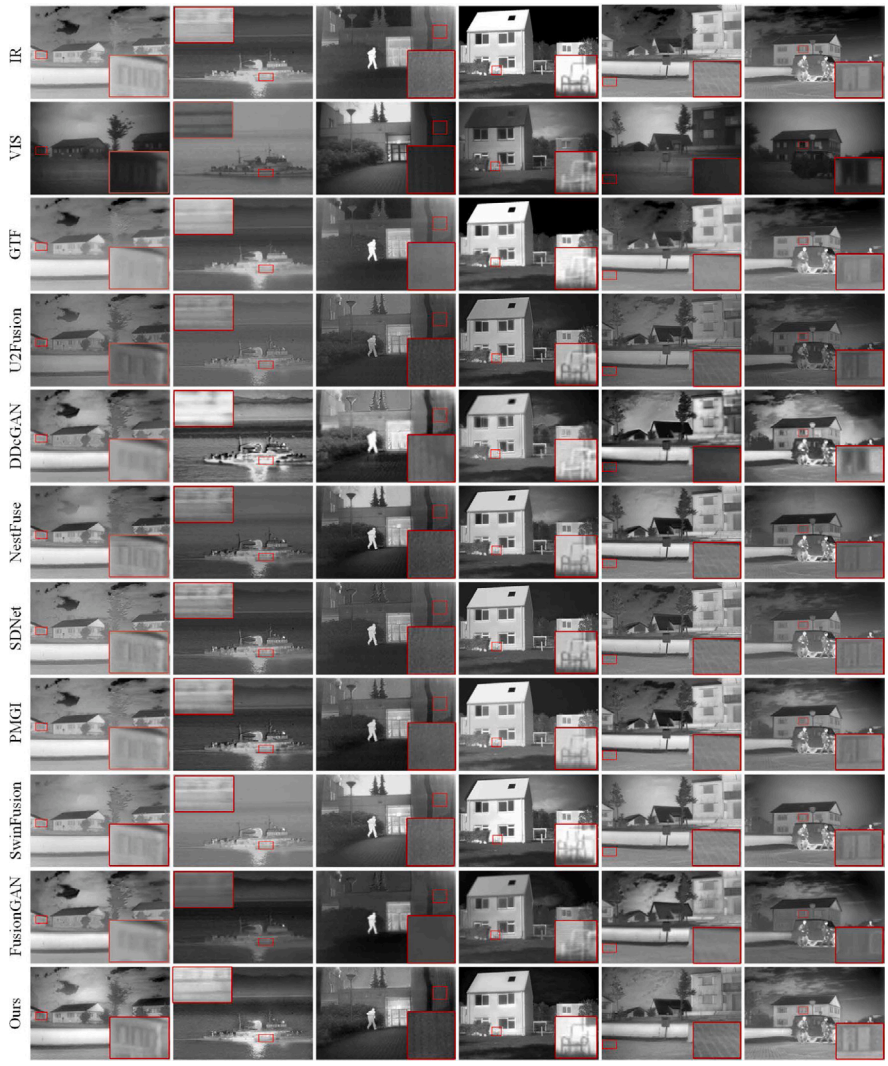

Fig. 9. Fusion results of different methods on source images without shifting pixels.

many existing methods, we select 16 pairs covering various scenarios from the TNO[4] and VOT2020-RGBT[5] datasets as the test set, and there are 3, 5, and 8 shifting pixels between any pair of images both vertically and horizontally. The selected 16 pairs are shown in Fig. 8.

4.2. Implementation details

In the training phase, we further randomly crop two patches for data argumentation. To obtain the infrared image with low-resolution which will be used in the alignment loss, the bilinear interpolation [68] is adopted to down-sample the source infrared image. Subsequently, infrared images with low-resolution are reconstructed whose scale factors are 0.5, 0.25 and 0.125, respectively. In our implementation, we set the batchsize to 8 and maximum of epochs to 120, and adopt the

[4] https://figshare.com/articles/dataset/TNO_Image_Fusion_Dataset/1008029

[5] https://www.votchallenge.net/vot2020/dataset

Fig. 10. Fusion results of different methods on source images with 5 shifting pixels on vertical and horizontal directions.

Adam [69] as the optimizer. For the learning rate, we primarily set it to 10^{-3}, which is decreased by following the scale factor 0.1 after each 40 epochs. Totally, we implement our model by Pytorch framework and train it on a GPU 3080TI. Besides, the Laplace operator is introduced to gain the gradient map.

4.3. Comparison methods and evaluation metrics

Here we make comparison with state-of-the-arts, including DDcGAN [36], U2Fusion [11], PMGI [13], NestFuse [23], FusionGAN [34], SDnet [41], SwinFusion [24] and GTF [70]. To quantitatively give the evaluation on these approaches, eight widely used metrics are adopted in this paper, which are Cross Entropy (CE) [71,72], Mutual Information (MI) [72–74], Edge based Similarity metric ($Q_{AB/F}$) [72,74,75], Chen-Blum Metric (Q_{CB}) [72,74,76], Structural Similarity Index Measure (SSIM) [72,77], Entropy (EN) [72,78], Chen-Varshney metric (Q_{CV}) [72,79] and Standard Deviation (SD) [72,80]. Specifically, CE evaluates the quality of the fusion results by measuring the information difference between the fused image and the source images. MI measures the correlation between fused image and source images. $Q_{AB/F}$ can assess the amount of the edges and details retained in the fused image from the source images. Considering the characteristics of the human visual system, Q_{CB} is used to appraise the magnitude of the information transformed from the sources images to the fused image from the perspective of local contrast. SSIM measures the similarity between fused images and source images from three aspects, including intensity, contrast and structure. EN reflects the degree of abundance of information contained in the fused images. Q_{CV} measures the local similarity between the fused image and source images based on the edge salience information. SD is used to evaluate the intensity difference of pixels in the fused image. Note that, except of CE and Q_{CV}, remaining metrics follow that the higher the values, the better the performances.

4.4. Comparison with image fusion methods

Fusion for Aligned Images: Although our fusion method is designed for images with translational misalignment, it is also adaptive for the case when the pair is completely aligned. Take the spatially aligned images as the inputs for different methods and the fused images are shown in Fig. 9. Despite the fact that some methods also achieve satisfactory performances, our presented model obtains much better visualization in textural details. The main reason is that our method takes more detailed information into account through the local feature refinement (LFR) strategy, which enhances the texture and prevents from distortion in the image reconstruction. To quantitatively measure different methods on aligned image fusion, Table 1 lists averaged results of aforementioned eight metrics for the 16 test pairs. Obviously, the proposed method enjoys dramatic performance enhancement compared with other methods.

Fusion for Misaligned Images: In this experiment, 16 translationally misaligned pairs are used as the inputs. Note that, they all encounter 3, 5 or 8 shifting pixels vertically and horizontally. To make a fair comparison, our misaligned training set is reused to train comparison methods and the predefined parameters are fixed according to their released codes.

Table 1
Objective evaluation results of different fusion methods on source images without shifting pixels. The bold and blue fonts indicate the optimal and suboptimal values, respectively.

Methods	CE	MI	$Q_{AB/F}$	Q_{CB}	Q_{CV}	SD	SSIM	EN
DDcGAN	0.9512	1.0183	0.4062	0.4936	1125.2534	**48.0880**	1.1086	**7.5008**
FusionGAN	2.0410	1.3367	0.2457	0.4353	1149.1628	23.7892	1.1703	6.8927
NestFuse	0.9099	1.9003	0.4997	0.5156	573.2640	41.3400	1.3048	7.0247
PMGI	0.9783	1.2311	0.4150	0.4885	610.6832	32.1324	1.2511	6.7459
SDNet	1.0763	0.9955	0.4290	0.5077	902.2495	25.3094	1.2227	6.8686
U2Fusion	0.5912	1.0557	0.4419	0.5452	632.7985	22.5613	1.3079	6.4650
GTF	0.7451	1.3456	0.3529	0.4345	1561.7065	25.4614	1.1705	6.3742
SwinFusion	0.4724	1.8912	0.4874	0.5592	533.1068	38.1374	1.3356	6.7799
Ours	**0.4514**	**1.9209**	**0.5072**	**0.5650**	**516.6418**	41.3804	**1.3414**	7.0563

Fig. 11. Fusion results of different methods on source images with different shifting pixels.

Fig. 10 displays the fused images when two source images encounter 5 shifting pixels. From the enlarged parts, it is easy to observe that all comparison methods meet heavy artifacts due to their miss-consideration on misaligned source images. By contrast, our proposed method is quite adaptive and robust for shifted images. As we can see in Fig. 10, there are not any artifacts compared with other methods. More importantly, this achievement is not dependent on image registration algorithms but directly obtained from feature extraction, which can be widely applied to other image fusion fields.

To further make comparison between our proposed methods and existing approaches, fused images with different shifting pixels are visualized in Fig. 11. As we can see, with the increasing number of shifting pixels, comparison methods suffer from much heavier artifacts. In contrast to them, there is not any inferior influence on results obtained by our proposed method. The change of quantitative evaluations are depicted in Fig. 12 when image pairs suffer from 3, 5 or 8 shifting pixels. Compared with existing state-of-the-arts, our proposed method achieves much superiority. Besides, we can see that quantitative values of other methods meet a large drop if two source images have larger numbers of shifting pixels. Differently, our method does not suffer from palpable performance degradation, further showing its effectiveness and robustness.

4.5. Comparison with image registration+fusion methods

As mentioned above, for existing image fusion methods, a straightforward way is to exploit the image registration first, although the large modality-gap does bring a large challenge. To substantiate this assumption, we utilize the state-of-the-art infrared–visible images registration algorithms HAPCG [22] and RIFT [54] first for image registration, followed by different fusion methods. Figs. 13 and 14 display the jointly registered and fused images when there are 5 shifting pixels. Of course, as shown on the first two rows in Fig. 13, if HAPCG achieves accurate registration, this two-step strategy does gain satisfactory performance. However, this case does not always meet success. As displayed on the third and forth rows in Fig. 13, the inaccurate registration results in artifacts and distortion in the fused images. Additionally, since two source images are almost complementary, it further increases the challenge for image registration. For instance, from the last two rows in Fig. 13, the registered infrared image is collapsed to a point. Obviously, the following fusion stage would miss the information from the infrared image. When RIFT is used as the registration method, as shown in Fig. 14, RIFT does not meet failure when aligning these three pairs of images. But there are still local misalignments in the last two pairs, introducing the artifacts in the fuse images subsequently. By contrast, our proposed method transforms the image registration pixel-by-pixel to high-level feature alignment, efficiently tackling this problem.

Table 2
Objective evaluation results of different registration+fusion methods on source images with 5 shifting pixels. The bold and blue fonts indicate the optimal and suboptimal values, respectively.

Configuration	CE	MI	$Q_{AB/F}$	Q_{CB}	Q_{CV}	SD	SSIM	EN
HAPCG+DDcGAN	1.1503	0.8874	0.1110	0.3702	3609.8687	**48.1705**	0.7819	5.8666
HAPCG+SDNet	1.3762	0.9160	0.1050	0.3547	3395.1710	34.6793	0.8089	5.1516
HAPCG+U2Fusion	1.3839	0.8512	0.2001	0.3845	2598.0969	32.0013	0.8904	6.0091
HAPCG+SwinFusion	1.0236	1.1756	0.2885	0.3464	3060.6668	39.0896	0.9039	5.0861
HAPCG+PMGI	1.5605	0.9850	0.1774	0.3278	3710.1109	38.4073	0.8014	4.9856
RIFT+DDcGAN	0.8904	1.1225	0.3036	0.4037	2503.6150	47.9555	0.8990	5.8056
RIFT+SDNet	0.7894	1.1856	0.2802	0.3288	3031.1002	36.9304	0.7492	5.7105
RIFT+U2Fusion	1.1902	1.0101	0.3175	0.4482	2008.4977	39.9093	0.8753	6.1921
RIFT+SwinFusion	0.9468	1.2072	0.3065	0.4104	1964.0638	40.8921	0.9470	5.9688
RIFT+PMGI	1.0577	0.9591	0.2750	0.3582	3083.5710	38.6720	0.7782	5.0616
Ours	**0.4517**	**1.9194**	**0.5073**	**0.5576**	**512.1710**	41.2566	**1.3389**	**7.0056**

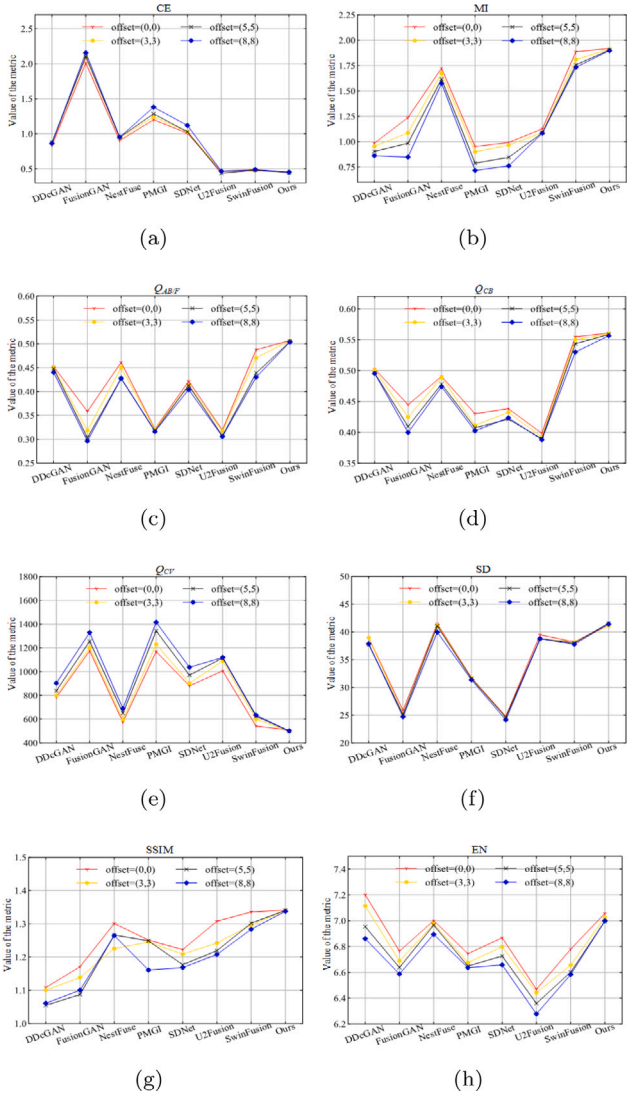

Fig. 12. Display of quality changes of different methods with different shifting pixels.

The associated values on CE, MI, $Q_{AB/F}$, Q_{CB}, Q_{CV}, SD, SSIM and EN metrics are displayed in Table 2. Due to occurrence of registration failure and texture warp caused by these registration process, the joint methods meet server performance regression. As we can observe, our approach obtains much better performance compared with these methods.

4.6. Experiments on real-world images

In the real-world scenario, for the sake of imaging and viewpoint difference, captured images from different sensors does encounter misalignment, which would be more complex than our synthetic data analyzed above. Therefore, we also conduct experiments on real-world images. Specifically, we randomly select 20 pairs from the CVC-14[6] dataset which consists of multispectral images for person detection. Experimental results are displayed in Fig. 15. Referring to DDcGAN and FusionGAN, since they fail to preserve detailed information, their computed results meet only slight artifacts. For remaining comparison methods, they do suffer from large performance degradation. In contrast to them, our proposed method not only removes the artifacts, but also enjoys abundant textural details in the fused images, which is more superior.

4.7. Ablation study

In this subsection, we conduct ablation studies to demonstrate the effectiveness of cross-modulation (CM) strategy, feature dynamic alignment module(FDAM) with feature alignment block (FAB) and dynamic alignment mechanism, complementary mask fusion (CMF) strategy, local feature refinement (LFR) strategy in the multi-grained feature refinement module (MGFRM), and pyramid feature fusion module (PFFM). As CM, FDAM, and LFR are introduced for feature alignment, we use the pairs which encounter 5 shifting pixels as the test images. By contrast, CMF and PFFM are proposed to retain more interested information for fused images, so we directly adopt the registered images as the inputs of the network for these two cases.

Effectiveness of CM: CM aims to extract the relationship between two source images so that the modulation parameters for two modalities are computed, which would be beneficial for following feature dynamic extraction in FDAM. When CM is removed from our proposed method, details of the fused image is shown in Fig. 16(b). Obviously, there is a slight artifact. Similarly, the corresponding quantitative evaluations in Table 3 also meet the drop, demonstrating the effectiveness of CM.

Effectiveness of FDAM: FDAM is proposed to learn dynamic kernels to align the infrared feature to the visible feature spatially. Fig. 16(c) shows the result obtained by our method when FDAM is removed. It is easy to observe that there is the artifact which is even heavier than that without CM. Additionally, we further analyze the influence of the number of FABs in FDAM. As visualized in Fig. 16(d), Fig. 16(e), and Fig. 16(f), we can observe that with the increasing number of FABs, artifacts in the fused images are gradually alleviated. Particularly, when 2 FABs are applied, fused images already meet our requirement. Furthermore, quantitative metrics in Table 3 also prove our analysis.

[6] adas.cvc.uab.es/elektra/enigma-portfolio/cvc-14-visible-fir-day-night-pedestrian-sequence-dataset/

Fig. 13. Fusion results of the registration+fusion methods and our method on source images with 5 shifting pixels. HAPCG [22] is used as registration method.

Fig. 14. Fusion results of the registration+fusion methods and our method on source images with 5 shifting pixels. RIFT [54] is used as registration method.

To further validate the effectiveness of the dynamic alignment mechanism, we remove the DKG block, and use the traditional convolution layers with same kernel size to take the place of the dynamic convolutional layers in FABs, which means the parameters are updated along with the training process and fixed at test phase. The comparative results are shown in Fig. 17. We can see that the artifacts emerge when the shifting pixels are more than 3 if we use the fixed convolutional kernels.

Effectiveness of CMF: To enjoy more valuable and complementary information from source images for the fused image, CMF is designed. To show its superiority, we replace it with the concatenation strategy followed by a 1×1 convolutional layer. As visualized in Fig. 18, the

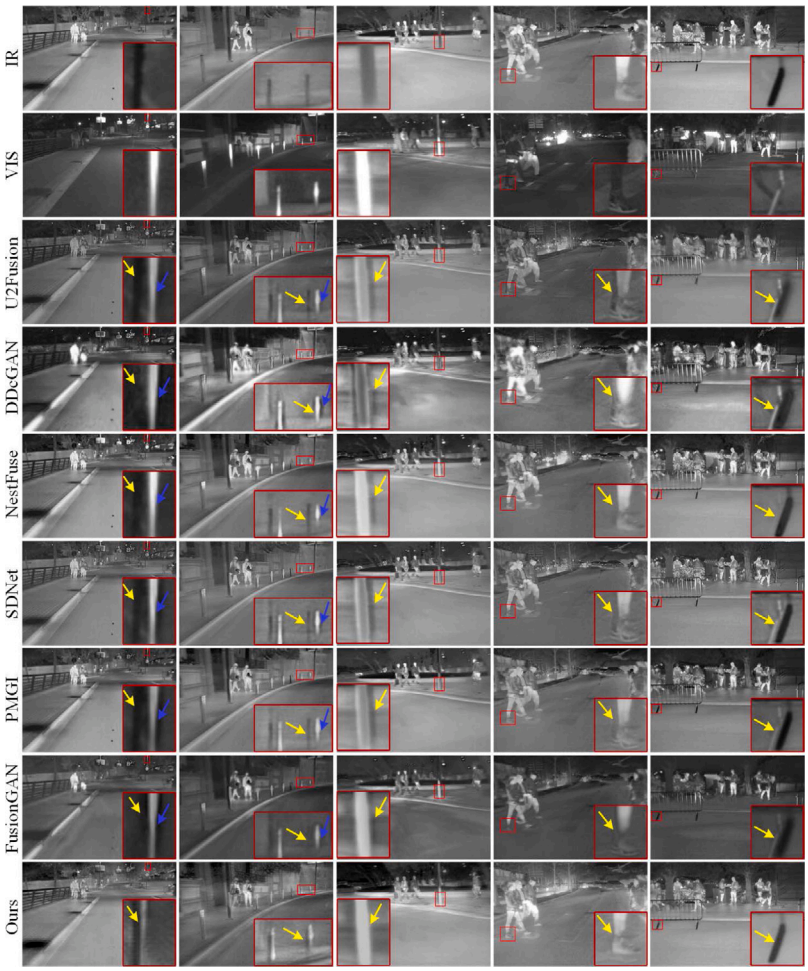

Fig. 15. Fusion results of different fusion methods on real-world images.

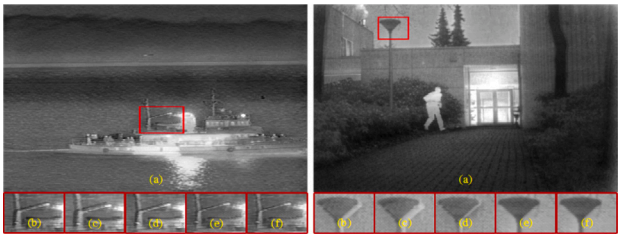

Fig. 16. Effectiveness validation of CM and FDAM. (a) are the fusion results of our overall network, (b) are the local areas of the fusion results of our network without CM, (c) are the local areas of the fusion results of our network without FDAM, (d)–(f) are the local areas of the fusion results of our network with 1, 2, 3 FABs, respectively.

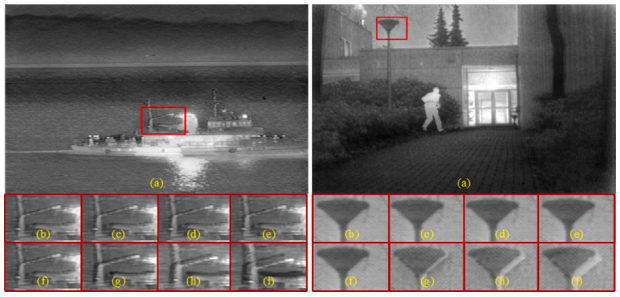

Fig. 17. Effectiveness validation of dynamic alignment mechanism. (a) are the fusion results of the network without DKG when fusing aligned source images, (b)–(e) are the local areas of the fusion results of our network when the shifting pixels are 0, 3, 5, 8, respectively. (f)–(i) are the local areas of the fusion results of our network without DKG when the shifting pixels are 0, 3, 5, 8, respectively.

modified version results in inferior visualization, especially in some detailed textures. So do the results in Table 3. Thus, our introduced CMF contributes to preserving more detailed information.

Effectiveness of PFFM: In our proposed approach, the pyramid-like architecture is exploited to fuse multi-scale features, so that more valuable information can be preserved. Similarly, here we also replace PFFM with deconvolutional layers + concatenation + 1 × 1 convolutional layer. Fig. 18 and Table 3 both indicate the significance of PFFM.

Effectiveness of LFR: LFR is composed of two branches: one is for texture enhancement, and another is for artifact removal. As displayed in Fig. 19(b,c,d), if two branches or one of them in LFR are deleted, fused images suffer from blur edges or the slight artifact. Fortunately, thanks to the embedding of LFR, the aforementioned problems are well addressed.

4.8. Parameter analysis

In our objective function, three hyper-parameters α, β, and λ should be predefined, making trade-off among the pixel loss l_{pixel}, the perceptual loss l_{per}, and the alignment loss l_{align}. Since l_{pixel} and l_{per} focus on preserving as much as information from the source images but do not take the alignment into account, we take the registered images as inputs

Table 3
Analysis of the effectiveness of different functional modules. 'RFIVF' denotes the proposed method. The bold and blue fonts indicate the optimal and suboptimal values, respectively.

Configuration	CE	MI	$Q_{AB/F}$	Q_{CB}	Q_{CV}	SD	SSIM	EN
w/o CM	0.4734	1.8935	0.4936	0.4993	543.2961	38.9755	1.2819	6.8606
w/o FDAM	0.4784	1.7186	0.4609	0.5088	631.1710	36.3593	1.1289	6.7515
w/o DKG	0.4702	1.8108	0.4691	0.5122	598.0649	38.9103	1.2190	6.9001
1 FAB	0.4568	1.8472	0.4925	0.5377	564.1710	38.8493	1.2339	6.8906
2 FABs	**0.4517**	**1.9194**	**0.5073**	**0.5576**	**512.1710**	**41.2566**	**1.3389**	**7.0056**
3 FABs	0.4548	1.9001	0.4961	0.5405	525.6926	40.5702	1.3062	6.9705
RFIVF	**0.4514**	**1.9209**	**0.5072**	**0.5650**	**516.6418**	**41.3804**	**1.3414**	**7.0563**
w/o CMF	0.4734	1.8935	0.4936	0.5431	523.2961	36.9755	1.2705	6.8616
w/o PFFM	0.4684	1.9086	0.4609	0.5588	568.1710	40.3550	1.2759	6.1710
w/o LFR	0.4925	1.7302	0.4582	0.5103	597.2831	38.6218	1.2045	6.5906
w/o upper branch	0.4703	1.8046	0.4719	0.5289	563.8392	38.7932	1.3055	6.8614
w/o lower branch	0.4594	1.8592	0.4828	0.5307	541.3794	39.8920	1.2959	6.8851

Table 4
Objective evaluation results of our method under different α. The bold and blue fonts indicate the optimal and suboptimal values, respectively.

α	CE	MI	$Q_{AB/F}$	Q_{CB}	Q_{CV}	SD	SSIM	EN
0.40	0.4740	1.8863	0.4954	0.5536	562.3928	39.3820	1.2795	6.7862
0.45	0.4578	1.9051	0.4910	0.5427	539.2792	40.3379	1.3028	6.8672
0.50	**0.4514**	**1.9209**	**0.5072**	**0.5650**	**516.6418**	**41.3804**	**1.3414**	**7.0563**
0.55	0.4596	1.9158	0.4899	0.5546	559.7480	41.0366	1.2829	6.9804
0.60	0.4702	1.8979	0.4942	0.5403	601.7822	38.8183	1.1806	6.6062

Fig. 18. Effectiveness validation of CMF and PFFM. From left to right: fusion results of our network without CMF, fusion results of our network without PFFM, and fusion results of our overall network.

Fig. 19. Effectiveness validation of LFR. (a) are the fusion results of our overall network, (b) are the local area of the fusion results of our network without LFR, (c) are the local area of the fusion results of our network without the upper branch in LFR, (d) are the local area of the fusion results of our network without the lower branch in LFR, (e) is the local area of the fusion results of our network with LFR.

to measure the effectiveness of these two loss functions. By contrast, we regard pairs suffering from 5 shifting pixels as the inputs for l_{align} measurement.

Analysis on α: To measure the susceptibility of α, we fix β and λ to 0.6 and 0.3, separately. Fig. 20 shows the visualizations of fused images under different α. When α is small, the contrast between objects and the background is also small as less infrared information is used. With the increase of α, the contrast tends to be more obvious while the detailed information, e.g., edges, become blurry. Thus, in this paper we set α to 0.5 empirically. Table 4 further substantiates our selection quantitatively.

Analysis on β: At this part, α and λ are set to 0.5 and 0.3 respectively. For β, we also empirically find that $\beta = 0.6$ is both adaptive for thermal information and textures, as shown in Fig. 21 and Table 5. Therefore, we set β to 0.6 in our experiments.

Analysis on λ: In the same way, the hyper-parameters α and β are separately set to 0.5 and 0.6. Fig. 22 and Table 6 demonstrate the importance of l_{align} under different λ. When λ is set to 0, there are noticeable artifacts around boundaries. Fortunately, when this parameter rises to 0.3, both visualization and quantitative values enjoy a remarkable improvement, indicating the significance of l_{align}. Thus, we set λ to 0.3 in our experiments.

4.9. Further discussion

Here we further conduct experiments on image pairs which suffer from 9, 11 and 13 shifting pixels, as depicted in Fig. 23. We can see that there do exist slight artifacts in the images fused by our method. In detail, with the rise of shifting pixels number, artifacts become more obvious. The main reason is that if source images encounter too heavy miss-registration, it is too challenging for the model to accurately learn the spatial relationship between them, resulting in performance degradation. In the real-world applications, the visible camera and the infrared camera have been well registered from their industry. However, due the external force, e.g., transportation, the two cameras would be slightly moved but the shifting pixels are statistically less than 9. Thus, our proposed method is adaptive for real-world applications.

5. Conclusion

In this paper, a novel method is proposed for infrared–visible image fusion which is robust to the translational misalignment between source images. Particularly, taking the visible modality as the reference, the modulation parameters and dynamic convolutional kernels are jointly learned according to the spatial relationship between two source images, so that their modalities are aligned in the high-level feature space. Followed by the attention, refinement, and a pyramid architecture, two different kinds of features are efficiently fused and reconstructed to an

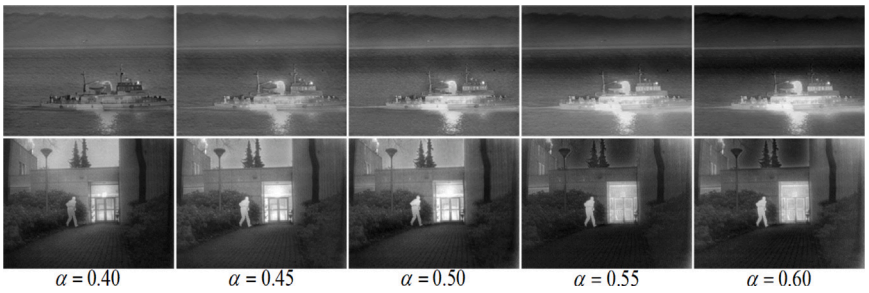

Fig. 20. Fusion results under different α.

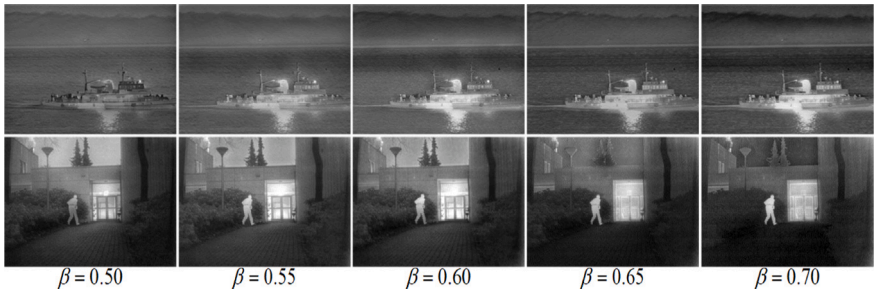

Fig. 21. Fusion results under different β.

Table 5
Objective evaluation results of our method under different β. The bold and blue fonts indicate the optimal and suboptimal values, respectively.

β	CE	MI	$Q_{AB/F}$	Q_{CB}	Q_{CV}	SD	SSIM	EN
0.50	0.4602	1.8934	0.4895	0.5486	558.3629	39.9378	1.2872	6.6207
0.55	0.4678	1.8802	0.4877	0.5493	531.2672	39.2389	1.3217	6.9279
0.60	**0.4514**	**1.9209**	**0.5072**	**0.5650**	**516.6418**	**41.3804**	**1.3414**	**7.0563**
0.65	0.4691	1.9101	0.4973	0.5551	552.1910	40.6526	1.3007	6.7792
0.70	0.4573	1.8958	0.4846	0.5271	572.6812	39.3871	1.2818	6.6720

Table 6
Objective evaluation results of our method under different λ. The bold and blue fonts indicate the optimal and suboptimal values, respectively.

λ	CE	MI	$Q_{AB/F}$	Q_{CB}	Q_{CV}	SD	SSIM	EN
0	0.4802	1.8894	0.4895	0.5296	558.3892	39.2798	1.2819	6.8072
0.1	0.4778	1.9011	0.4873	0.5372	537.2856	39.1739	1.2901	6.9518
0.2	0.4684	1.8936	0.4993	0.5339	521.2618	40.0217	1.3218	6.8157
0.3	**0.4517**	**1.9194**	**0.5073**	**0.5576**	**512.1710**	**41.2566**	**1.3389**	**7.0056**
0.4	0.4584	1.9028	0.4931	0.5463	523.7311	40.3836	1.3009	6.9277
0.5	0.4673	1.8947	0.4826	0.5315	530.2382	39.0371	1.2869	6.8877

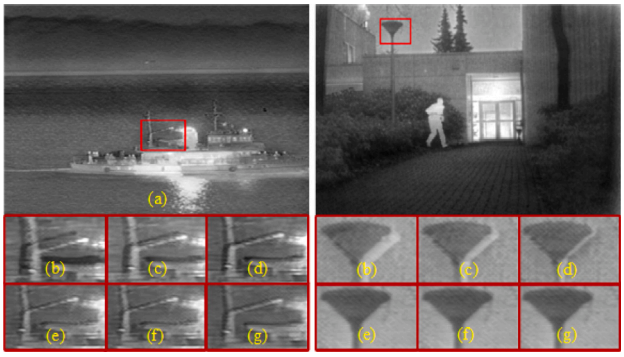

Fig. 22. Fusion results under different λ.

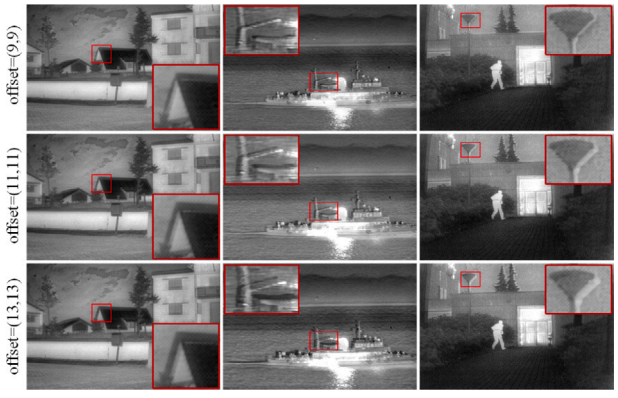

Fig. 23. Fusion results on source images with more than 9 shifting pixels.

image without any artifact but enjoying both thermal and textural details. Experiments conducted on both synthetic and real-world datasets substantiate the superiority of our proposed method compared with the state-of-the-arts. To the best of our knowledge, this is the first work which is completely free from image registration, providing a novel idea for real-world multi-modal image fusion.

CRediT authorship contribution statement

Huafeng Li: Ideas, Methodology, Experimental design, Formal analysis. **Junzhi Zhao:** Ideas, Methodology, Software, Validation, Data curation. **Jinxing Li:** Writing – original draft, Writing – review & editing. **Zhengtao Yu:** Supervision, Writing – review & editing. **Guangming Lu:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported in part by the NSFC, China fund (62272133, 62276120, 61966021, 62161015), the Shenzhen Colleges and Universities Stable Support Program, China (GXWD20220811170100001) and Shenzhen Science and Technology Program, China (RCBS20200714114910193).

References

[1] S. Gao, Y. Cheng, Y. Zhao, Method of visual and infrared fusion for moving object detection, Opt. Lett. 38 (11) (2013) 1981–1983.
[2] H. Kaur, D. Koundal, V. Kadyan, Image fusion techniques: A survey, Arch. Comput. Methods Eng. 28 (7) (2021) 4425–4447.
[3] H. Zhang, H. Xu, X. Tian, J. Jiang, J. Ma, Image fusion meets deep learning: A survey and perspective, Inf. Fusion 76 (2021) 323–336.
[4] J. Ma, Y. Ma, C. Li, Infrared and visible image fusion methods and applications: A survey, Inf. Fusion 45 (2019) 153–178.
[5] X. Jin, Q. Jiang, S. Yao, D. Zhou, R. Nie, J. Hai, K. He, A survey of infrared and visual image fusion methods, Infrared Phys. Technol. 85 (2017) 478–501.
[6] S. Li, X. Kang, L. Fang, J. Hu, H. Yin, Pixel-level image fusion: A survey of the state of the art, Inf. Fusion 33 (2017) 100–112.
[7] Y. Feng, H. Lu, J. Bai, L. Cao, H. Yin, Fully convolutional network-based infrared and visible image fusion, Multimedia Tools Appl. 79 (21) (2020) 15001–15014.
[8] H. Li, Y. Cen, Y. Liu, X. Chen, 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. 30 (2021) 4070–4083.
[9] Y. Liu, X. Chen, J. Cheng, H. Peng, Z. Wang, Infrared and visible image fusion with convolutional neural networks, Int. J. Wavelets Multiresol. Inf. Proc. 16 (03) (2018) 1850018.
[10] H. Xu, J. Ma, J. Yuan, Z. Le, W. Liu, RFNet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 19679–19688.
[11] H. Xu, J. Ma, J. Jiang, X. Guo, H. Ling, U2Fusion: A unified unsupervised image fusion network, IEEE Trans. Pattern Anal. Mach. Intell. 44 (1) (2020) 502–518.
[12] H. Xu, J. Ma, Z. Le, J. Jiang, X. Guo, FusionDN: A unified densely connected network for image fusion, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 12484–12491.
[13] H. Zhang, H. Xu, Y. Xiao, X. Guo, 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. 12797–12804.
[14] J. Ma, L. Tang, M. Xu, H. Zhang, G. Xiao, STDFusionNet: An infrared and visible image fusion network based on salient target detection, IEEE Trans. Instrum. Meas. 70 (2021) 5009513.
[15] J. Ma, J. Jiang, H. Zhou, J. Zhao, X. Guo, Guided locality preserving feature matching for remote sensing image registration, IEEE Trans. Geosci. Remote Sens. 56 (8) (2018) 4435–4447.
[16] J. Ma, J. Jiang, C. Liu, Y. Li, Feature guided Gaussian mixture model with semi-supervised EM and local geometric constraint for retinal image registration, Inform. Sci. 417 (2017) 128–142.
[17] J. Ma, J. Zhao, Y. Ma, J. Tian, Non-rigid visible and infrared face registration via regularized Gaussian fields criterion, Pattern Recognit. 48 (3) (2015) 772–784.
[18] G. Wang, Z. Wang, Y. Chen, Q. Zhou, W. Zhao, Context-aware Gaussian fields for non-rigid point set registration, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 5811–5819.
[19] J. Ma, J. Zhao, A.L. Yuille, Non-rigid point set registration by preserving global and local structures, IEEE Trans. Image Process. 25 (1) (2015) 53–64.
[20] D. Wang, J. Liu, X. Fan, R. Liu, Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration, 2022, arXiv preprint arXiv:2205.11876.
[21] L. Tang, Y. Deng, Y. Ma, J. Huang, J. Ma, SuperFusion: A versatile image registration and fusion network with semantic awarenes, IEEE/CAA J. Autom. Sin. 9 (12) (2022) 2121–2137.
[22] Y. Yao, Y. Zhang, Y. Wan, X. Liu, Heterologous images matching considering anisotropic weighted moment and absolute phase orientation, Geomat. Inf. Sci. Wuhan Univ. 46 (11) (2021) 1727–1736.
[23] H. Li, X.-J. Wu, T. Durrani, NestFuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models, IEEE Trans. Instrum. Meas. 69 (12) (2020) 9645–9656.
[24] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, Y. Ma, SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer, IEEE/CAA J. Autom. Sin. 9 (7) (2022) 1200–1217.
[25] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.
[26] Y. Liu, X. Chen, H. Peng, Z. Wang, Multi-focus image fusion with a deep convolutional neural network, Inf. Fusion 36 (2017) 191–207.
[27] H. Li, X.-J. Wu, DenseFuse: A fusion approach to infrared and visible images, IEEE Trans. Image Process. 28 (5) (2018) 2614–2623.
[28] F. Lahoud, S. Süsstrunk, Fast and efficient zero-learning image fusion, 2019, arXiv preprint arXiv:1905.03590.
[29] R. Liu, J. Liu, Z. Jiang, X. Fan, Z. Luo, A bilevel integrated model with data-driven layer ensemble for multi-modality image fusion, IEEE Trans. Image Process. 30 (2020) 1261–1274.
[30] L. Jian, X. Yang, Z. Liu, G. Jeon, M. Gao, D. Chisholm, A symmetric encoder-decoder with residual block for infrared and visible image fusion, 2019, arXiv preprint arXiv:1905.11447.
[31] H. Jung, Y. Kim, H. Jang, N. Ha, K. Sohn, Unsupervised deep image fusion with structure tensor representations, IEEE Trans. Image Process. 29 (2020) 3845–3858.
[32] H. Li, X.-J. Wu, J. Kittler, RFN-Nest: An end-to-end residual fusion network for infrared and visible images, Inf. Fusion 73 (2021) 72–86.
[33] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Adv. Neural Inf. Process. Syst. 27 (2014).
[34] J. Ma, W. Yu, P. Liang, C. Li, J. Jiang, FusionGAN: A generative adversarial network for infrared and visible image fusion, Inf. Fusion 48 (2019) 11–26.
[35] J. Ma, P. Liang, W. Yu, C. Chen, X. Guo, J. Wu, J. Jiang, Infrared and visible image fusion via detail preserving adversarial learning, Inf. Fusion 54 (2020) 85–98.
[36] J. Ma, H. Xu, J. Jiang, X. Mei, X.-P. Zhang, DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion, IEEE Trans. Image Process. 29 (2020) 4980–4995.
[37] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, 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.
[38] L. Tang, J. Yuan, 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 82 (2022) 28–42.
[39] L. Tang, J. Yuan, H. Zhang, X. Jiang, J. Ma, PIAFusion: A progressive infrared and visible image fusion network based on illumination aware, Inf. Fusion 83 (2022) 79–92.
[40] Y. Sun, B. Cao, P. Zhu, 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.
[41] H. Zhang, J. Ma, SDNet: A versatile squeeze-and-decomposition network for real-time image fusion, Int. J. Comput. Vis. 129 (10) (2021) 2761–2785.
[42] Z. Le, J. Huang, H. Xu, F. Fan, Y. Ma, X. Mei, J. Ma, UIFGAN: An unsupervised continual-learning generative adversarial network for unified image fusion, Inf. Fusion 88 (2022) 305–318.
[43] P. Liang, J. Jiang, X. Liu, J. Ma, Fusion from decomposition: A self-supervised decomposition approach for image fusion, in: European Conference on Computer Vision, Springer, 2022, pp. 719–735.
[44] J. Ma, X. Jiang, A. Fan, J. Jiang, J. Yan, Image matching from handcrafted to deep features: A survey, Int. J. Comput. Vis. 129 (1) (2021) 23–79.

[45] Y. Liu, L. Wang, J. Cheng, C. Li, X. Chen, Multi-focus image fusion: A survey of the state of the art, Inf. Fusion 64 (2020) 71–91.
[46] G. Lv, S.W. Teng, G. Lu, Enhancing SIFT-based image registration performance by building and selecting highly discriminating descriptors, Pattern Recognit. Lett. 84 (2016) 156–162.
[47] H. Bay, T. Tuytelaars, L.V. Gool, Surf: Speeded up robust features, in: European Conference on Computer Vision, Springer, 2006, pp. 404–417.
[48] L. Huang, C. Chen, H. Shen, B. He, Adaptive registration algorithm of color images based on SURF, Measurement 66 (2015) 118–124.
[49] Y. Ke, R. Sukthankar, PCA-SIFT: A more distinctive representation for local image descriptors, in: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2, 2004 CVPR 2004, IEEE, 2004, p. II.
[50] W. Ma, Z. Wen, Y. Wu, L. Jiao, M. Gong, Y. Zheng, L. Liu, Remote sensing image registration with modified SIFT and enhanced feature matching, IEEE Geosci. Remote Sens. Lett. 14 (1) (2016) 3–7.
[51] E. Rosten, T. Drummond, Machine learning for high-speed corner detection, in: European Conference on Computer Vision, Springer, 2006, pp. 430–443.
[52] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, ORB: An efficient alternative to SIFT or SURF, in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 2564–2571.
[53] J. Ma, J. Zhao, J. Jiang, H. Zhou, X. Guo, Locality preserving matching, Int. J. Comput. Vis. 127 (5) (2019) 512–531.
[54] J. Li, Q. Hu, M. Ai, RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform, IEEE Trans. Image Process. 29 (2019) 3296–3310.
[55] X. Yan, Y. Zhang, D. Zhang, N. Hou, B. Zhang, Registration of multimodal remote sensing images using transfer optimization, IEEE Geosci. Remote Sens. Lett. 17 (12) (2020) 2060–2064.
[56] X. Xie, Y. Zhang, X. Ling, X. Wang, A novel extended phase correlation algorithm based on Log-Gabor filtering for multimodal remote sensing image registration, Int. J. Remote Sens. 40 (14) (2019) 5429–5453.
[57] K.M. Yi, E. Trulls, Y. Ono, V. Lepetit, M. Salzmann, P. Fua, Learning to find good correspondences, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2666–2674.
[58] J. Ma, X. Jiang, J. Jiang, J. Zhao, X. Guo, LMR: Learning a two-class classifier for mismatch removal, IEEE Trans. Image Process. 28 (8) (2019) 4045–4059.
[59] L. Shen, J. Zhu, C. Fan, X. Huang, T. Jin, A novel Affine covariant feature mismatch removal for feature matching, IEEE Trans. Geosci. Remote Sens. 60 (2021) 1–13.
[60] Y. Wu, J.-W. Liu, C.-Z. Zhu, Z.-F. Bai, Q.-G. Miao, W.-P. Ma, M.-G. Gong, Computational intelligence in remote sensing image registration: A survey, Int. J. Autom. Comput. 18 (1) (2021) 1–17.
[61] J. Li, D. Fan, L. Yang, S. Gu, G. Lu, Y. Xu, D. Zhang, Layer-output guided complementary attention learning for image defocus blur detection, IEEE Trans. Image Process. 30 (2021) 3748–3763.
[62] X. Wang, K. Yu, C. Dong, C.C. Loy, Recovering realistic texture in image super-resolution by deep spatial feature transform, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 606–615.
[63] J. Liang, G. Sun, K. Zhang, L. Van Gool, R. Timofte, Mutual affine network for spatially variant kernel estimation in blind image super-resolution, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, CVPR, 2021, pp. 4096–4105.
[64] H. Song, W. Xu, D. Liu, B. Liu, Q. Liu, D.N. Metaxas, Multi-stage feature fusion network for video super-resolution, IEEE Trans. Image Process. 30 (2021) 2923–2934.
[65] J. Johnson, A. Alahi, F. Li, Perceptual losses for real-time style transfer and super-resolution, in: European Conference on Computer Vision, Springer, 2016, pp. 694–711.
[66] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.
[67] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.
[68] P. Smith, Bilinear interpolation of digital images, Ultramicroscopy 6 (2) (1981) 201–204.
[69] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.
[70] J. Ma, C. Chen, C. Li, J. Huang, Infrared and visible image fusion via gradient transfer and total variation minimization, Inf. Fusion 31 (2016) 100–109.
[71] D. Bulanon, T. Burks, V. Alchanatis, Image fusion of visible and thermal images for fruit detection, Biosyst. Eng. 103 (1) (2009) 12–22.
[72] X. Zhang, P. Ye, G. Xiao, VIFB: A visible and infrared image fusion benchmark, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPRW, 2020, pp. 104–105.
[73] G. Qu, D. Zhang, P. Yan, Information measure for performance of image fusion, Electron. Lett. 38 (7) (2002) 1.
[74] Z. Liu, E. Blasch, Z. Xue, J. Zhao, R. Laganiere, W. Wu, Objective assessment of multiresolution image fusion algorithms for context enhancement in night vision: A comparative study, IEEE Trans. Pattern Anal. Mach. Intell. 34 (1) (2011) 94–109.
[75] C.S. Xydeas, V. Petrovic, et al., Objective image fusion performance measure, Electron. Lett. 36 (4) (2000) 308–309.
[76] Y. Chen, R.S. Blum, A new automated quality assessment algorithm for image fusion, Image Vis. Comput. 27 (10) (2009) 1421–1432.
[77] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: From error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[78] J.W. Roberts, J.A. Van Aardt, F.B. Ahmed, Assessment of image fusion procedures using entropy, image quality, and multispectral classification, J. Appl. Remote Sens. 2 (1) (2008) 023522.
[79] H. Chen, P.K. Varshney, A human perception inspired quality metric for image fusion based on regional information, Inf. Fusion 8 (2) (2007) 193–207.
[80] Y.-J. Rao, In-fibre Bragg grating sensors, Meas. Sci. Technol. 8 (4) (1997) 355.