Unsupervised Misaligned Infrared and Visible Image Fusion via Cross-Modality Image Generation and Registration

Di Wang[1] , **Jinyuan Liu**[1] , **Xin Fan**[2] and **Risheng Liu**[2*]

[1]School of Software Technology, Dalian University of Technology, China
[2]International School of Information Science Engineering, Dalian University of Technology, China
diwang1211@mail.dlut.edu.cn, atlantis918@hotmail.com, {xin.fan, rsliu}@dlut.edu.cn

Abstract

Recent learning-based image fusion methods have marked numerous progress in pre-registered multi-modality data, but suffered serious ghosts dealing with misaligned multi-modality data, due to the spatial deformation and the difficulty narrowing cross-modality discrepancy. To overcome the obstacles, in this paper, we present a robust cross-modality generation-registration paradigm for unsupervised misaligned infrared and visible image fusion (IVIF). Specifically, we propose a Cross-modality Perceptual Style Transfer Network (CP-STN) to generate a pseudo infrared image taking a visible image as input. Benefiting from the favorable geometry preservation ability of the CP-STN, the generated pseudo infrared image embraces a sharp structure, which is more conducive to transforming cross-modality image alignment into mono-modality registration coupled with the structure-sensitive of the infrared image. In this case, we introduce a Multi-level Refinement Registration Network (MRRN) to predict the displacement vector field between distorted and pseudo infrared images and reconstruct registered infrared image under the mono-modality setting. Moreover, to better fuse the registered infrared images and visible images, we present a feature Interaction Fusion Module (IFM) to adaptively select more meaningful features for fusion in the Dual-path Interaction Fusion Network (DIFN). Extensive experimental results suggest that the proposed method performs superior capability on misaligned cross-modality image fusion.

1 Introduction

Image fusion is to extract the complementary information from different modality images and aggregate them to generate richer and more meaningful feature representations. Typically, Infrared and Visible Image Fusion (IVIF) enjoys the merit and presents conduciveness to practical applications such as autonomous driving and video surveillance.

Most existing IVIF methods [Liu *et al.*, 2020b; 2021e; 2022b] are designed for the hand-crafted pre-registered images, which are very labor-intensive and time-sensitive, despite numerous progress. And, they are sensitive to the discrepancy of intensity of the misaligned infrared and visible images, resulting in serious ghosting artifacts on fused images, once slight offset and deformation exist. The intrinsic reason is that the large cross-modality variation between infrared and visible images makes it impractical to straightly bridge the domain gap between them in a shared feature space. Coupled with the absence of cross-modality similarity constraints, few attempts have ever been made to fuse the misaligned infrared and visible images. The primary obstacle is cross-modality image alignment.

More often than not, existing widely-used image alignment methods [Ilg *et al.*, 2017; Balakrishnan *et al.*, 2018] perform pixel-wise and feature-level alignment by explicitly estimating the deformation field between the distorted image and its reference. However, they are worked under only a single modality setting [Tang *et al.*, 2020] due to their high dependence on the similarity of distribution and appearance on either synthetic or real data with neighborhood reference. Moreover, the designs of the similarity measures [Tang *et al.*, 2022] to optimize the cross-modality alignment process have been proven to be quite challenging. These obstacles gave rise to the development of cross-modality medical image translation [Yang *et al.*, 2020; Liu *et al.*, 2020a; 2021d] and cross-modality Re-ID tasks [Wang *et al.*, 2019a; 2019b]. The former leverages the cGAN [Isola *et al.*, 2017] to perform NeuroImage-to-NeuroImage translation, while the latter learns the cross-modality match in pedestrian images of a person across disjoint camera views by an RGB-to-IR (infrared) translation. A recent study [Arar *et al.*, 2020] presents a multi-modality image registration method by training an image-to-image translation network on the two input modalities. The basic idea of the above-mentioned approaches is using the image-to-image translation to implement cross-modality transformation to narrow the large variation between different modalities.

Motivated by this idea, we propose a specialized cross-

---

*Corresponding author

modality generation-registration paradigm (CGRP) for unsupervised misaligned infrared and visible image fusion. An inherent characteristic of infrared images is the emphasis on sharp geometry structures over texture details. Generating a pseudo infrared image preserving sharp structure information is more conducive to implementing mono-modality registration between it with the corresponding distorted infrared image. To better preserve geometry structure during generating pseudo infrared images from visible images, we proposed a Cross-modality Perceptual Style Transfer Network (CPSTN), which inherits the basic cycle consistent learning manner of the CycleGAN [Zhu *et al.*, 2017], while developing a perceptual style transfer constraint and cross regularizations across two-cycle learning path to further guide the CPSTN to generate sharper structures. Without them, the generation capability of our CPSTN will degrade to that in [Arar *et al.*, 2020]. Doing this lays the foundation for the mono-modality registration of infrared images. We utilize a Multi-level Refinement Registration Network (MRRN) to predict the deformation field from coarse to fine between distorted and pseudo infrared images and reconstruct the registered infrared image. In this case, we further perform the fusion of the registered infrared image and visible image. To enable the fusion network to focus more on faithful texture details, while avoiding feature smoothing caused by some unsophisticated fusion rules (e.g., concatenation, weighted average), we develop an Interaction Fusion Module (IFM) to adaptively select meaningful features from infrared and visible images to achieve fused results with sharp textures. We evaluate the proposed method on manually processed unaligned datasets and demonstrate its strengths through comprehensive analysis. The main contributions are summarized as follows:

- We propose a highly-robust unsupervised infrared and visible image fusion framework, which is more focused on mitigating ghosting artifacts caused by the fusion of misaligned image pairs compared with learning-based fusion methods specialized for pre-registered images.

- Considering the difficulty of the cross-modality image alignment, we exploit a specialized cross-modality generation-registration paradigm to bridge the large discrepancy between modalities, thus achieving effective infrared and visible image alignment.

- An interaction fusion module is developed to adaptively fuse multi-modality features, which avoids feature smoothing caused by unsophisticated fusion rules and emphasizes faithful texture details.

The extensive experimental results demonstrate that the proposed method performs superior capability on misaligned cross-modality image fusion. The code will be available at https://github.com/wdhudiekou/UMF-CMGR.

2 Method

2.1 Motivation

Different imaging pipelines and heat dissipation within the sensor result in misalignment between observed infrared and visible images, presenting as shift and deformation. By observing, it is found that straightly fusing misaligned infrared and visible images often suffer serious ghosts. Inspired

by [Wang *et al.*, 2019a; 2019b] that reduce cross-modality variation via image-to-image translation, and catering to the inherent characteristic that infrared image "emphasizes structure over texture", we propose a specialized cross-modality generation-registration paradigm to reduce spatial offsets and alleviate the ghost during misaligned infrared and visible image fusion.

2.2 Cross-modality Perceptual Style Transfer

The first part of the proposed cross-modality generation-registration paradigm (CGRP) is image translation. Considering that infrared images are prone to distortion due to thermal radiation, we proposed a Cross-modality Perceptual Style Transfer Network (CPSTN) to translate the visible image I_{vis} to its infrared counterpart $I_{\tilde{ir}}$. Together, a pseudo infrared image pair $(I_{ir}, I_{\tilde{ir}})$ is formed to provide a unified representation. As shown in Figure 1, CPSTN is a UNet-like generator, and its bottom obtains 9 resnet blocks. Different from CycleGAN [Zhu *et al.*, 2017], we tend to design a specific learning strategy controlled by a perceptual style transfer constraint and establish the inter-path correlation between the two-cycle generative routes to further optimize [Liu *et al.*, 2022c; 2021c] the generation of sharp structure of the pseudo infrared image. The pseudo infrared image is generated by $I_{\tilde{ir}} = \mathcal{T}_\theta(I_{ir})$, where $\mathcal{T}_\theta$ denotes the CSPTN with network parameter θ. This optimization process is illustrated in the left dotted box in Figure 1. Note that the $G(A)$ corresponds to our CPSTN, and the perceptual style transfer constraint and the regularization of inter-path correlation are explained in Section 2.5.

2.3 Multi-level Refinement Registration

Thanks to the reduction of cross-modality discrepancy by CPSTN, the registration of infrared images under the mono-modality settings becomes the other core part of the CGRP. As shown in Figure 1, we exploit a Multi-level Refinement Registration Network (MMRN) to predict deformation field between distorted and pseudo infrared images and reconstruct the registered infrared image. The MRRN consists of a shared multi-level feature extractor (SM-FE), two Coarse-to-Fine deformation field estimation (C2F-DFE) modules, and a resampler layer. In each C2F-DFE, a coarse DFE module $\mathcal{M}_C$ and a refined DFE module $\mathcal{M}_R$ are included. Then, the coarse deformation field is first predicted as

$$\varphi_c^k = \mathcal{M}_C\left(\mathcal{F}^k(I_{\tilde{ir}}, I_{ir})\right), \qquad (1)$$

the refined deformation field is estimated by

$$\varphi_r^k = \mathcal{M}_R\left(\varphi_c^k\right) \oplus \varphi_c^k, \qquad (2)$$

where $\mathcal{F}^k$ denotes the k-th level of the SM-FM. Suppose the SM-FE contains K levels, when $k = K$, the final deformation field $-\phi = \varphi_r^K$ is estimated. Finally, we use the resampler layer similar to the STN [Jaderberg *et al.*, 2015] to reconstruct the registered infrared image by

$$I_{ir}^{reg} = I_{ir} \circ (-\phi), \qquad (3)$$

the operator $\circ$ denotes the spatial transform for registration.

2.4 Dual-path Interaction Fusion

To fuse registered infrared image I_{ir}^{reg} and visible image I_{vis}, we present a dual-path interaction fusion network (DIFN). It
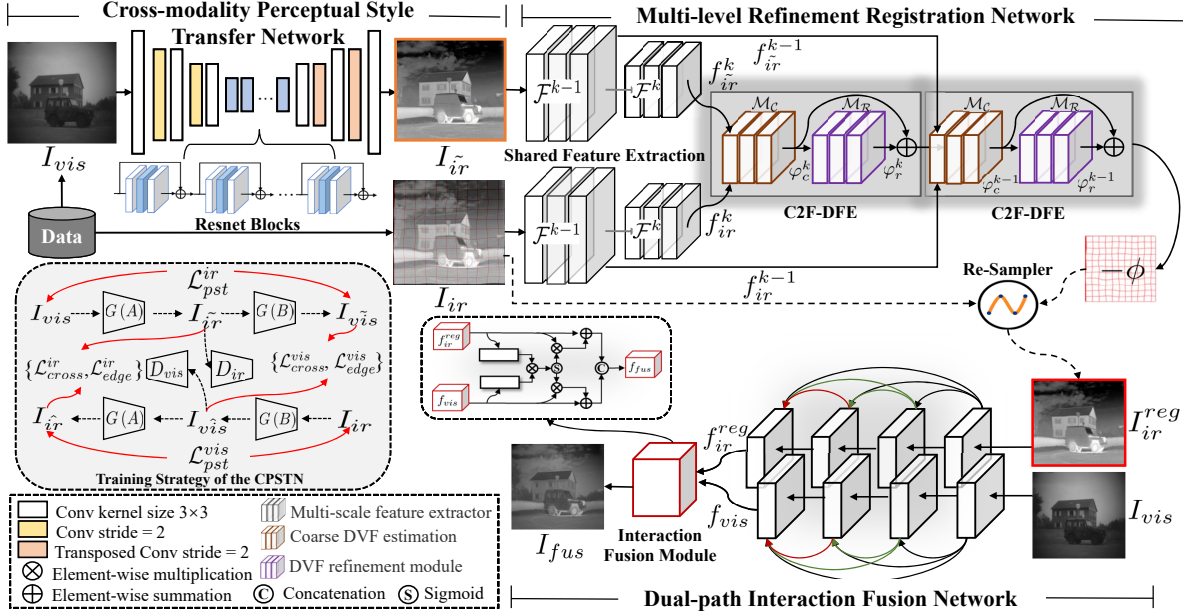
Figure 1: The workflow of the proposed unsupervised cross-modality fusion network for misaligned infrared and visible images. The network mainly consists of three subnetworks, which are Cross-modality Perceptual Style Transfer Network, Multi-level Refinement Registration Network, and Dual-path Interaction Fusion Network. Taking misaligned infrared and visible images as inputs, our method performs the above three subnetworks in sequence to obtain the final fused image.

is composed of a dual-path feature extraction module and a feature interaction fusion module (IFM). Here, the structure of the dual-path feature extraction module inherits the residual dense network [Zhang *et al.*, 2021b], and the extracted features are presented by

$$f_{ir}^{reg}, f_{vis} = \mathcal{M}_{\theta_E}\left(I_{ir}^{reg}, I_{vis}\right), \quad (4)$$

where $\mathcal{M}_{\theta_E}$ is the feature extraction module, and θ_E is its parameter.

Interaction Fusion Module. We utilize the IFM to adaptively select features from infrared and visible images to fuse, shown in small dashed box of Figure 1. To focus on more significant information, the feature responses are recalibrated by

$$\mathcal{A}_{tt} = \mathcal{S}\left(Conv_{1\times 1}\left(f_{ir}^{reg}\right) \otimes Conv_{1\times 1}\left(f_{vis}\right)\right), \quad (5)$$

and the infrared and visible features are activated as

$$\begin{aligned} f_{ir}^{Att} &= f_{ir}^{reg} \otimes \left(1 + \mathcal{A}_{tt}\right), \\ f_{vis}^{Att} &= f_{vis} \otimes \left(1 + \mathcal{A}_{tt}\right), \end{aligned} \quad (6)$$

then we obtain the final fused image by

$$I_{fus} = Conv_{3\times 3}\left(Concat\left(f_{ir}^{Att}, f_{vi}^{Att}\right)\right), \quad (7)$$

where $\mathcal{S}$ is the Sigmoid function, and $\otimes$ denotes the element-wise multiplication operation.

2.5 Loss Functions

Perceptual style transfer loss. To generate more realistic pseudo infrared images, we introduce a perceptual style transfer (PST) loss to control the cycle consistency of CPSTN. The PST loss consists of two terms known as perceptual loss $\mathcal{L}_{pcp}$ and style loss $\mathcal{L}_{sty}$. First, the $\mathcal{L}_{pcp}$ is defined as

$$\begin{aligned} \mathcal{L}_{pcp}^{\psi_j} &= \|\psi_j(I_{vis}) - \psi_j\left(G_B\left(G_A(I_{vis})\right)\right)\|^2 \\ &+ \|\psi_j(I_{ir}) - \psi_j\left(G_A\left(G_B(I_{ir})\right)\right)\|^2, \end{aligned} \quad (8)$$

where ψ_j is the j-th layer of the VGG-19 [Simonyan and Zisserman, 2015] model and $j \in [2, 7, 12, 21, 30]$ along with weight $\omega \in [\frac{1}{32}, \frac{1}{16}, \frac{1}{8}, 1, 1]$. These features are also used to compute the $\mathcal{L}_{sty}$, which is defined as

$$\begin{aligned} \mathcal{L}_{sty}^{\psi_j} &= \omega_j \|\mathcal{G}_{\psi_j}(I_{vis}) - \mathcal{G}_{\psi_j}\left(G_B\left(G_A(I_{vis})\right)\right)\|^2 \\ &+ \omega_j \|\mathcal{G}_{\psi_j}(I_{ir}) - \mathcal{G}_{\psi_j}\left(G_A\left(G_B(I_{ir})\right)\right)\|^2, \end{aligned} \quad (9)$$

where $\mathcal{G}$ is Gram matrix [Sajjadi *et al.*, 2017], used to combat "checkerboard" artifacts. The overall PST loss is

$$\mathcal{L}_{pst} = \lambda_p \mathcal{L}_{pcp} + \lambda_s \mathcal{L}_{sty}, \quad (10)$$

where λ_p is set to 1.0, λ_s is set to 100.

Cross regularization loss. Inventively, we present a cross regularization between two-cycle paths during training CPSTN to establish the inter-path correlation. It contains both content term $\mathcal{L}_{con}$ and edge term $\mathcal{L}_{edge}$, which are defined as

$$\begin{aligned} \mathcal{L}_{con} &= \|I_{i\tilde{r}} - I_{i\hat{r}}\|_1 + \|I_{v\tilde{i}s} - I_{v\hat{i}s}\|_1, \\ \mathcal{L}_{edge} &= \|\nabla I_{i\tilde{r}} - \nabla I_{i\hat{r}}\|_{char} + \|\nabla I_{v\tilde{i}s} - \nabla I_{v\hat{i}s}\|_{char}, \end{aligned} \quad (11)$$

where ∇ denotes laplacian gradient operator. We compute $\mathcal{L}_{edge}$ using the Charbonnier Loss [Lai *et al.*, 2019]. The overall cross regularization is computed by

$$\mathcal{L}_{cross} = \lambda_c \mathcal{L}_{con} + \lambda_e \mathcal{L}_{edge}, \quad (12)$$

where, $\lambda_c = \lambda_e = 8.0$.

Registration loss. We adopt a bidirectional similarity loss to constrain the registration between distorted and pseudo infrared images in feature space, which is defined as

$$\mathcal{L}_{sim}^{bi} = \|\psi_j(I_{ir}^{reg}) - \psi_j(I_{i\tilde{r}})\| + \lambda_{rev}\|\psi_j(\phi \circ I_{i\tilde{r}}) - \psi_j(I_{ir})\|_1, \quad (13)$$

the first term is forward, while the second term is backward with weight $\lambda_{rev} = 0.2$, in which the reverse deformation

field ϕ is used to distort the pseudo infrared image $I_{\tilde{ir}}$ and make it approach to the distorted input I_{ir}. To ensure a smooth deformation field, we define a smooth loss as

$$\mathcal{L}_{\text{smooth}} = \|\nabla \phi\|_1. \tag{14}$$

And then, the overall registration loss is computed by

$$\mathcal{L}_{\text{reg}} = \mathcal{L}_{sim} + \lambda_{sm}\mathcal{L}_{smooth}, \tag{15}$$

where λ_{sm} is set to 10 in our work.

Fusion loss. In the fuison phase, we utilize MS-SSIM loss function $\mathcal{L}_{ssim}^{ms}$ to maintain sharper intensity distribution of the fused image, which is defined as

$$\mathcal{L}_{ssim}^{ms} = (1 - SSIM(I_{fus}, I_{ir}^{reg})) + (1 - SSIM(I_{fus}, I_{vis})). \tag{16}$$

To encourage the restoration of texture details, we model the gradient distribution and develop a joint gradient loss as

$$\mathcal{L}_{JGrad} = \|max(\nabla I_{ir}^{reg}, \nabla I_{vis}), \nabla I_{fus}\|_1. \tag{17}$$

For clearer textures, the gradient of the fused image is forced to approach the maximum value between the infrared and visible image gradients. In addition, to retain the saliency targets from the two images, we leverage self-visual saliency maps inspried by [Ghosh *et al.*, 2019; Liu *et al.*, 2021b; 2022a] to build a new loss as

$$\boldsymbol{\omega}_{ir} = \boldsymbol{S}_{f_{ir}}/(\boldsymbol{S}_{f_{ir}} - \boldsymbol{S}_{f_{vis}}), \boldsymbol{\omega}_{vis} = 1 - \boldsymbol{\omega}_{ir},$$
$$\mathcal{L}_{svs} = \|(\boldsymbol{\omega}_{ir} \otimes I_{ir}^{reg} + \boldsymbol{\omega}_{vis} \otimes I_{vis}), I_{fus}\|_1, \tag{18}$$

where $\boldsymbol{S}$ denotes saliency matrix, $\boldsymbol{\omega}_{vis}$ and $\boldsymbol{\omega}_{ir}$ denote the weight maps for infrared and visible images, respectively. The overall fusion loss is computed by

$$\mathcal{L}_{fus} = \lambda_{ssim}\mathcal{L}_{ssim}^{ms} + \lambda_{JG}\mathcal{L}_{JGrad} + \lambda_{svs}\mathcal{L}_{svs}, \tag{19}$$

where λ_{ssim}, λ_{JG}, and λ_{svs} are set to 1.0, 20.0, and 5.0.

Overall loss. We train our network by minimizing the following overall loss function

$$\mathcal{L}_{total} = \mathcal{L}_{pst} + \mathcal{L}_{cross} + \mathcal{L}_{GAN} + \mathcal{L}_{reg} + \mathcal{L}_{fus}. \tag{20}$$

Note that the $\mathcal{L}_{GAN}$ inherits that of CycleGAN to discriminate whether pseudo infrared image is real or fake.

3 Experiments

3.1 Dataset and Implement Details

Datasets. Our misalignment cross-modality image fusion are conducted on two widely-used datasets: TNO[1], and RoadScene[2], which are pre-registered. To meet the requirement of misaligned image pairs, we first generate several deformation fields by performing different degrees of affine and elastic transformations and then apply them to infrared images to obtain distorted images. We use all 221 images from RoadScene dataset as training samples. Due to the limited amount of the TNO dataset, we only use it as a testing set. In addition, we randomly select 55% images from RoadScene dataset for testing since our method is full-unsupervised.

Implement Details. The code of our method is implemented using the PyTorch framework with an NVIDIA 2080Ti GPU. We randomly select 8 image patches of size 256×256 to form a batch. The Adam optimizer ($\beta_1 = 0.9$, and $\beta_2 = 0.999$) is used to optimize our model. The initial learning rate is set to 0.001 and remains unchanged throughout the training phase that goes through 300 epochs.

---

[1]http://figshare.com/articles/TNO_Image_Fusion_Dataset/1008029.
[2]https://github.com/hanna-xu/RoadScene.

3.2 Evaluation in IR-VIS Image Registration

We evaluate the middle registered results using three common metrics including MSE, Mutual Information (MI) [Mahapatra *et al.*, 2018], and Normalized Cross Correlation (NCC) [Cao *et al.*, 2018]. Considering that multi-modality image alignment methods are extremely rare, for a fair comparison with ours, we choose two typical image alignment algorithms based on deformation field, which are FlowNet [Ilg *et al.*, 2017] and VoxelMorph [Balakrishnan *et al.*, 2018]. Note that, FlowNet only predicts the deformation field for input image pairs, so the spatial transform layer (STN) [Jaderberg *et al.*, 2015] is employed to generate registered images. As reported by Table 1, using directly these two algorithms to perform cross-modality alignment of the IR-VIS images yields negligible improvement, even worse than the misaligned inputs. In contrast, the proposed CGRP gains about **0.10** and **0.23** improvments in MI on TNO and RoadScene datasets, respectively. It is demonstrated that the proposed CGRP is more effective than existing popular image alignment methods.

Methods	TNO			RoadScene		
	MSE↓	NCC↑	MI↑	MSE↓	NCC↑	MI↑
Misaligned Input	0.007	0.876	1.558	0.01	0.894	1.602
FlowNet + STN	0.029	0.636	1.095	0.009	0.910	1.549
FlowNet + STN + CPST	0.007	0.893	1.465	0.005	0.949	1.744
VoxelMorph	0.007	0.880	1.545	0.008	0.914	1.589
VoxelMorph + CPST	0.005	0.919	1.573	0.006	0.941	1.689
Ours CGRP	**0.004**	**0.926**	**1.648**	**0.004**	**0.963**	**1.833**

Table 1: Quantitative comparisons of cross-modality image registration on the TNO and Roadscene datasets. The best results are presented with **red** bold font.

3.3 Comparing with SOTA in IVIF

We perform quantitative and qualitative evaluations for the proposed method against several state-of-the-art IVIF methods, including DENSE [Li and Wu, 2019], DIDFuse [Zhao *et al.*, 2020], FGAN [Ma *et al.*, 2019], GAN-FM [Zhang *et al.*, 2021a], MFEIF [Liu *et al.*, 2021a], RFN [Li *et al.*, 2021], and U2F [Xu *et al.*, 2022], equipped with FlowNet+STN and VoxelMorph, two typical image alignment algorithms used to cross-modality image registration in our work. For fair comparisons, we use the same testing samples (121 and 24 images from RoadScene and TNO datasets) to evaluate them.

Quantitative evaluation. We report the quantitative fusion results on TNO and RoadScene datasets in Table 2. It can be seen that our methods numerically outperforms existing IVIF methods by large margins and ranks first across all three metrics, including Cross Correlation (CC), Visual Information Fidelity (VIF) [Han *et al.*, 2013], and Structural Similarity Index (SSIM) [Wang *et al.*, 2004]. Especially, compared with GAN-FM, a recent research, our method gains **0.33** and **0.46** improvements in VIF on TNO and RoadScene datasets, which demonstrate the superiority of the proposed method.

Qualitative evaluation. Due to space limitation, Figure 2 only presents qualitative comparisons using FlowNet+STN as the basic registration model. By observing the local enlarged areas, we can find that prior works fail to achieve desirable alignment and ghost elimination. In contrast, the pro-

	DENSE	DIDFuse	FGAN	GAN-FM	MFEIF	PMGI	RFN	U2F	Ours
①: Aligned Method: (FlowNet+STN) / VoxelMorph ②: Dataset: RoadScene									
CC↑	0.589/**0.597**	0.573/0.582	0.502/0.582	0.565/0.575	0.588/0.596	0.530/0.538	0.587/0.593	0.558/0.569	**0.621**
VIF↑	0.488/0.521	0.445/0.469	0.493/0.513	0.564/0.597	0.619/**0.659**	0.408/0.447	0.563/0.590	0.430/0.464	**0.895**
SSIM↑	0.311/0.355	0.301/0.335	0.255/0.281	0.334/0.368	0.342/**0.380**	0.266/0.314	0.319/0.347	0.297/0.343	**0.507**
①: Aligned Method: (FlowNet+STN) / VoxelMorph ②: Dataset: TNO									
CC↑	0.390/0.463	0.380/0.452	0.315/0.400	0.350/0.419	0.381/0.451	0.363/0.447	0.392/**0.465**	0.378/0.456	**0.481**
VIF↑	0.675/0.742	0.545/0.597	0.588/0.654	0.556/0.627	0.701/**0.780**	0.518/0.606	0.671/0.728	0.535/0.601	**1.016**
SSIM	0.318/**0.382**	0.266/0.328	0.236/0.294	0.295/0.372	0.297/0.379	0.277/0.363	0.300/0.349	0.299/0.372	**0.473**

Table 2: The quantitative comparisons of the IVIF results with eight state-of-the-art fusion methods on two common datasets when using FlowNet+STN and VoxelMorph algorithms as the basic registration models.

	DENSE	FGAN	DIDFuse	U2F	PMGI	RFN	MFEIF	GAN-FM	Ours
Param.	0.925	**0.074**	0.261	0.659	**0.042**	10.94	0.158	10.21	0.80
Time	0.124	0.251	0.055	0.123	0.182	0.239	**0.045**	0.334	**0.024**

Table 3: Efficiency analysis of our joint registration-fusion model against state-of-the-art fusion methods.

posed method shows favorable alignment and fusion capability while preserving sharp structures.

Model efficiency. To examine model efficiency, we report the statistics of Params(M) and RunTime(s) in Table 3. We conduct the measures for a paired images with size of 64×64 on a single 2080Ti GPU. Noted that, our results are obtained by jointly measuring registration and fusion networks, while other methods only have a single fusion network. The results show that our method is competitive at least in running time.

3.4 Ablation Studies

CPSTN. We plug the CPSTN into FlowNet and VoxelMorph as their improved versions, then adopt them to conduct cross-modality IR-VIS image registration. As shown in Table 1, their quantitative results improve a large margin in collaboration with CPSTN than the original versions. Accordingly, the visual comparisons provided in Figure 3 suggest the effectiveness of the CPSTN. We observe that the registered results generated by the FlowNet model equipped with the CPSTN eliminate evident distortion. Besides, we verify the effectiveness of the CPSTN from the perspective of image fusion. As quantified in Table 4, improvements of **0.14** and **0.22** in VIF are obtained using the CPSTN on TNO and RoadScene datasets, respectively. The corresponding qualitative comparisons in Figure 4(b) and (c) show that the CPSTN contributes to favorable fusion results with negligible ghosts for misaligned infrared and visible images. The above results comprehensively reveal the effectiveness of CPSTN from registration and fusion perspectives.

MRRN. Suppose that in the absence of CPSTN, we investigate the effect of MRRN on misaligned IVIF tasks. Note that MRRN performs cross-modality image registration taking the paired infrared and visible images as inputs under the same settings. As reported in Table 4, we compare the results of the first two experiments on the TNO and RoadScene datasets, which show that using MRRN to implement cross-modality registration also improves the performance of IVIF to a certain extent, compared with straightly fusing the misaligned infrared and visible images.

Datasets	CPSTN	MRRN	DIFN	CC↑	SSIM↑	VIF↑
TNO	✗	✗	✓	0.438	0.369	0.732
	✗	✓	✓	$0.476_{\uparrow 0.038}$	$0.418_{\uparrow 0.049}$	$0.876_{\uparrow 0.144}$
	✓	✓	✓	$\mathbf{0.481}_{\uparrow 0.044}$	$\mathbf{0.473}_{\uparrow 0.103}$	$\mathbf{1.016}_{\uparrow 0.374}$
Road	✗	✗	✓	0.563	0.367	0.543
	✗	✓	✓	$0.601_{\uparrow 0.038}$	$0.422_{\uparrow 0.055}$	$0.673_{\uparrow 0.130}$
	✓	✓	✓	$\mathbf{0.621}_{\uparrow 0.085}$	$\mathbf{0.507}_{\uparrow 0.140}$	$\mathbf{0.895}_{\uparrow 0.352}$

Table 4: Ablation studies of the CPSTN and MRRN on TNO and RoadScene datasets.

Loss Functions in CPSTN. We analyze the effect of $\mathcal{L}_{pst}$ and $\mathcal{L}_{cross}$ used in the CPSTN. As shown in Figure 5, without using the constraints from $\mathcal{L}_{pst}$ and $\mathcal{L}_{cross}$, the generated pseudo infrared image suffers serious structural degradation (Figure 5(c)) compared with the reference image (Figure 5(a)). Using $\mathcal{L}_{cross}$, the model preserves general structural information, while subtle structures are not maintained well enough and "checkerboard artifacts" are introduced obviously (Figure 5(d)). In contrast, the pseudo infrared image generated by our model (Figure 5(e)) has a sharper geometry structure, which caters to the common sense that infrared image "emphasizes structure over texture".

Interaction Fusion Module. We replace the IFM in the DIFN by concatenation operation to examine its effectiveness for the IVIF. As observed in Figure 6, the model without using the IFM tends to generate smooth textures with low contrast, while the result using the IFM has richer and sharper textures with positively high contrast. It is indicated that adaptively selecting features to be fused is more effective for improving fused image quality.

3.5 Additional Analysis

To verify the generalization of the proposed CPSTN, we plug it into two typical image alignment algorithms (i.e., FlowNet+STN and VoxelMorph) implement mono-modality image registration between distorted and pseudo infrared images, and further to implement fusion of the registered infrared image and visible image on eight existing IVIF methods. As shown in Figure 7, the first box of each method is the result of the model without using the CPSTN, and the second box is that of the model using the CPSTN. It can be observed that: i) These alignment models equipped with the CPSTN can promote effectively the fusion performance of existing

VIS/IR   DENSE   DIDFuse   FGAN   GAN-FM   MFEIF   PMGI   RFN   U2F   Ours
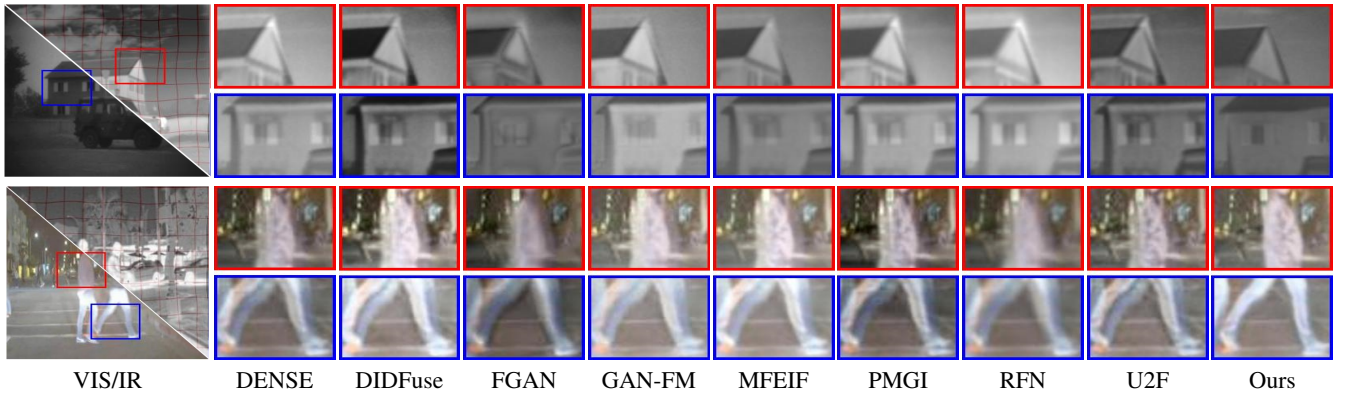
Figure 2: Visualization of the IVIF results from different methods on the TNO and RoadScene datasets when using the FlowNet+STN algorithm as the basic registration model.
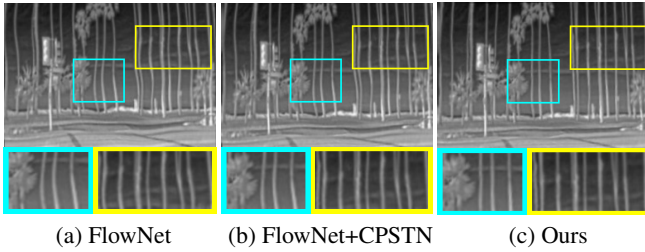

(a) FlowNet   (b) FlowNet+CPSTN   (c) Ours

Figure 3: Ablation analysis of the CPSTN from the registration perspective on RoadScene dataset.

(a) w/o $\mathcal{T},\mathcal{R}$ (b) w/o $\mathcal{T}$ (c) Ours
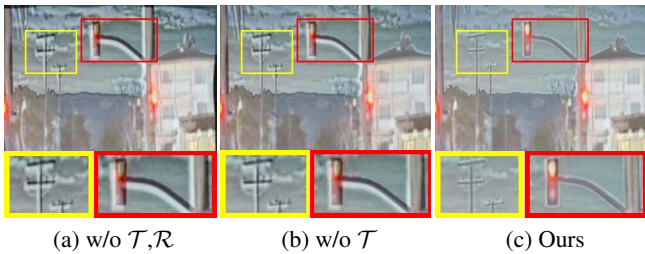
Figure 4: Ablation analysis of the CPSTN ($\mathcal{T}$) and MRRN ($\mathcal{R}$) on RoadScene dataset.

IVIF methods for misaligned images, which reveals the favorable generalization ability of the CPSTN for misaligned IVIF tasks. ii) Although conditioned by i), the proposed method still outperforms these improved IVIF methods on misaligned cross-modality images, which explains that the superiority of our overall framework relies on the close cooperation of each component.

4 Conclusion

In this work, we proposed a highly-robust unsupervised misaligned infrared and visible image fusion framework for mitigating ghosts of the fused images. We exploited a generation-registration paradigm to simplify the cross-modality image alignment into a mono-modality registration. A feature interaction fusion module was developed to adaptively select meaningful features from infrared and visible images to fuse, avoiding feature smoothing and emphasizing faithful textures. Extensive experimental results demonstrated the superior capability of our method on misaligned cross-modality image fusion. Significantly, our generation-
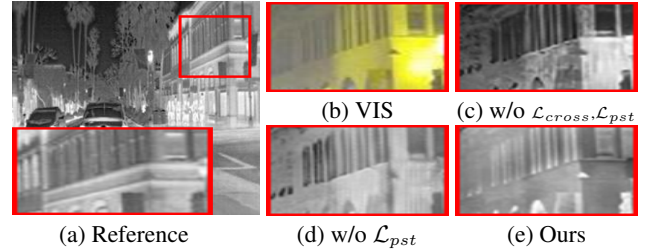

(a) Reference   (d) w/o $\mathcal{L}_{pst}$ (e) Ours

(b) VIS   (c) w/o $\mathcal{L}_{cross},\mathcal{L}_{pst}$

Figure 5: Ablation analysis of loss functions in CPSTN on RoadScene dataset.

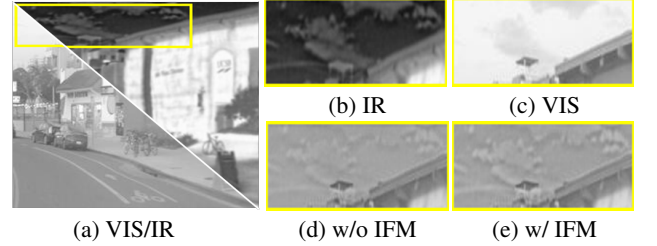(a) VIS/IR   (d) w/o IFM   (e) w/ IFM

(b) IR   (c) VIS

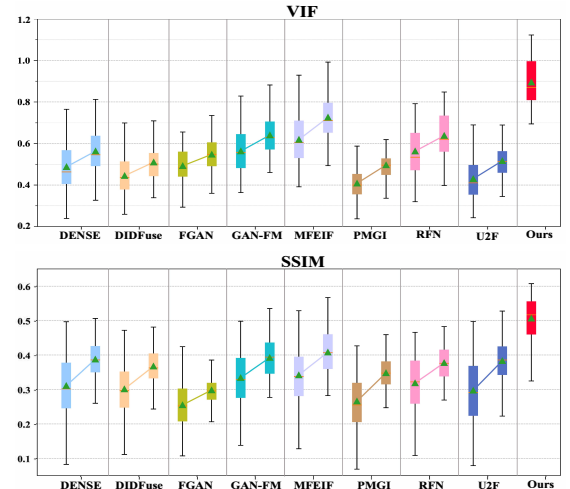Figure 6: Ablation analysis of the IFM on RoadScene dataset.

Figure 7: Generalization analysis of the proposed CPSTN.

registration paradigm can be well extended to existing IVIF methods to improve their fusion performance on misaligned cross-modality images.

Acknowledgments

References

[Arar *et al.*, 2020] Moab Arar, Yiftach Ginger, Dov Danon, Amit H. Bermano, and Daniel Cohen-Or. Unsupervised multi-modal image registration via geometry preserving image-to-image translation. In *CVPR*, pages 13407–13416, 2020.

[Balakrishnan *et al.*, 2018] Guha Balakrishnan, Amy Zhao, Mert R. Sabuncu, John V. Guttag, and Adrian V. Dalca. An unsupervised learning model for deformable medical image registration. In *CVPR*, pages 9252–9260, 2018.

[Cao *et al.*, 2018] Xiaohuan Cao, Jianhuan Yang, Li Wang, Zhong Xue, Qian Wang, and Dinggang Shen. Deep learning based inter-modality image registration supervised by intra-modality similarity. In *MLMI*, pages 55–63, 2018.

[Ghosh *et al.*, 2019] Sanjay Ghosh, Ruturaj G. Gavaskar, and Kunal N. Chaudhury. Saliency guided image detail enhancement. In *NCC*, pages 1–6, 2019.

[Han *et al.*, 2013] Yu Han, Yunze Cai, Yin Cao, and Xiaoming Xu. A new image fusion performance metric based on visual information fidelity. *Information Fusion*, 14(2):127–135, 2013.

[Ilg *et al.*, 2017] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, pages 1647–1655, 2017.

[Isola *et al.*, 2017] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 5967–5976, 2017.

[Jaderberg *et al.*, 2015] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *NeurIPS*, pages 2017–2025, 2015.

[Lai *et al.*, 2019] Weisheng Lai, Jiabin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Fast and accurate image super-resolution with deep laplacian pyramid networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(11):2599–2613, 2019.

[Li and Wu, 2019] Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2019.

[Li *et al.*, 2021] Hui Li, Xiao-Jun Wu, and Josef Kittler. Rfn-nest: An end-to-end residual fusion network for infrared and visible images. *Information Fusion*, 73:72–86, 2021.

[Liu *et al.*, 2020a] Risheng Liu, Zi Li, Yuxi Zhang, Xin Fan, and Zhongxuan Luo. Bi-level probabilistic feature learning for deformable image registration. In *IJCAI*, pages 723–730, 2020.

[Liu *et al.*, 2020b] Risheng Liu, Jinyuan Liu, Zhiying Jiang, Xin Fan, and Zhongxuan Luo. A bilevel integrated model with data-driven layer ensemble for multi-modality image fusion. *IEEE Transactions on Image Processing*, 30:1261–1274, 2020.

[Liu *et al.*, 2021a] Jinyuan Liu, Xin Fan, Ji Jiang, Risheng Liu, and Zhongxuan Luo. Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[Liu *et al.*, 2021b] Jinyuan Liu, Yuhui Wu, Zhanbo Huang, Risheng Liu, and Xin Fan. Smoa: Searching a modality-oriented architecture for infrared and visible image fusion. *IEEE Signal Processing Letters*, 28:1818–1822, 2021.

[Liu *et al.*, 2021c] Risheng Liu, Jiaxin Gao, Jin Zhang, Deyu Meng, and Zhouchen Lin. Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[Liu *et al.*, 2021d] Risheng Liu, Zi Li, Xin Fan, Chenying Zhao, Hao Huang, and Zhongxuan Luo. Learning deformable image registration from optimization: Perspective, modules, bilevel training and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[Liu *et al.*, 2021e] Risheng Liu, Zhu Liu, Jinyuan Liu, and Xin Fan. Searching a hierarchically aggregated fusion architecture for fast multi-modality image fusion. In *ACM MM*, pages 1600–1608, 2021.

[Liu *et al.*, 2022a] Jinyuan Liu, Xin Fan, Zhangbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *CVPR*, 2022.

[Liu *et al.*, 2022b] Jinyuan Liu, Jingjie Shang, Risheng Liu, and Xin Fan. Attention-guided global-local adversarial learning for detail-preserving multi-exposure image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[Liu *et al.*, 2022c] Risheng Liu, Pan Mu, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. A general descent aggregation framework for gradient-based bi-level optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[Ma *et al.*, 2019] Jiayi Ma, Wei Yu, Pengwei Liang, Chang Li, and Junjun Jiang. Fusiongan: A generative adversarial network for infrared and visible image fusion. *Information Fusion*, 48:11–26, 2019.

[Mahapatra *et al.*, 2018] Dwarikanath Mahapatra, Zongyuan Ge, Suman Sedai, and Rajib Chakravorty. Joint registration and segmentation of xray images using generative adversarial networks. In *MICCAI*, pages 73–80, 2018.

[Sajjadi *et al.*, 2017] Mehdi S. M. Sajjadi, Bernhard Schölkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *ICCV*, pages 4501–4510, 2017.

[Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[Tang *et al.*, 2020] Hao Tang, Zechao Li, Zhimao Peng, and Jinhui Tang. Blockmix: Meta regularization and self-calibrated inference for metric-based meta-learning. In *ACM MM*, pages 610–618, 2020.

[Tang *et al.*, 2022] Hao Tang, Chengcheng Yuan, Zechao Li, and Jinhui Tang. Learning attention-guided pyramidal features for few-shot fine-grained recognition. *Pattern Recognition*, 2022.

[Wang *et al.*, 2004] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[Wang *et al.*, 2019a] Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In *ICCV*, pages 3622–3631, 2019.

[Wang *et al.*, 2019b] Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shin'ichi Satoh. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *CVPR*, pages 618–626, 2019.

[Xu *et al.*, 2022] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 44(1):502–518, 2022.

[Yang *et al.*, 2020] Qianye Yang, Nannan Li, Zixu Zhao, Xingyu Fan, Eric I-Chao Chang, and Yan Xu. Mri cross-modality image-to-image translation. *Scientific reports*, 10(1):1–18, 2020.

[Zhang *et al.*, 2021a] Hao Zhang, Jiteng Yuan, Xin Tian, and Jiayi Ma. GAN-FM: infrared and visible image fusion using GAN with full-scale skip connection and dual markovian discriminators. *IEEE Transactions on Computational Imaging*, 7:1134–1147, 2021.

[Zhang *et al.*, 2021b] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(7):2480–2495, 2021.

[Zhao *et al.*, 2020] Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, Jiangshe Zhang, and Pengfei Li. Didfuse: Deep image decomposition for infrared and visible image fusion. In *IJCAI*, pages 970–976, 2020.

[Zhu *et al.*, 2017] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2242–2251, 2017.