DSTFuse: Enhancing Deblurring via Style Transfer for Visible and Infrared Image Fusion

Gang Pan
Tianjin University
pangang@tju.edu.cn

Yonglu Liu
Tianjin University
lyl991002@tju.edu.cn

Jinyuan Li
Tianjin University
jinyuanli@tju.edu.cn

Zhenjun Han
University of Chinese Academy of Sciences
hanzhj@ucas.ac.cn

Jiahao Wang
Tianjin University
wjhwtt@tju.edu.cn

Di Sun
Tianjin University of Science and Technology
dsun@tust.edu.cn

Abstract

Infrared and visible image fusion aims at obtaining fused images that keep advantages of source images, e.g., detailed textures and clear edge structures. To tackle the challenge in modeling features from visible image under motion blur and low light conditions, we propose a novel fusion framework, DSTFuse, which aims to leverage infrared image as the style image and enable it to perform style transfer on the visible image to efficiently eliminate motion blur. Specifically, DSTFuse contains a Cross-Modality Style Transfer Module (CST-module) that collect appropriate style information from the infrared image and guide the transformation of blurry objects into the corresponding style while preserve all other elements without alteration. The output of CST-module is integrated with the image with a multitude of visible features from another module and mapped into final image. Extensive experiments show that DSTFuse achieves promising results in infrared-visible image fusion task. And it is also shown that DSTFuse can boost the performance in downstream infrared-visible object detection. Code will be released at https://anonymous.4open.science/r/DSTFuse-0C1D.

1. Introduction

Image fusion is a fundamental image enhancement technique. It aims to combine images with distinct modality features into a image that retains the advantage of the source images [1, 34, 35, 49, 50, 54]. One prevalent application of image fusion is the infrared and visible image fusion (IVIF) [31, 39, 40, 42]. Proverbially, visible images can reflect the appearance and color information of objects, while infrared images provide thermal radiation information, characterized by a high contrast between the target and its surroundings. By integrating the complementary information from both visible and infrared images, IVIF generates a fused image that that overcomes the limitations of visible images under environmental constraints and the lack of detail in infrared images. Therefore, IVIF has broad applications across various fields such as military [19], security [9], and medical image processing [50].

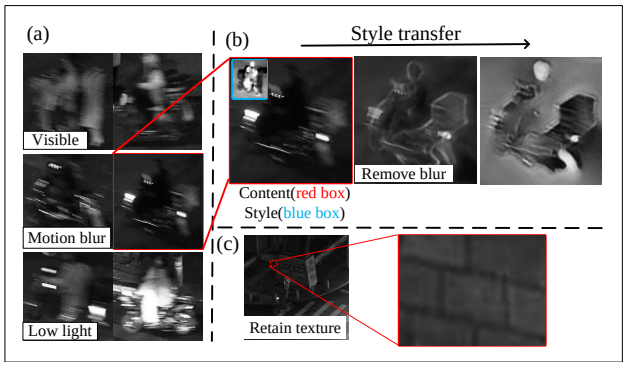

Figure 1. The effect of style transfer on suboptimal visible data for IVIF tasks. (a) The suboptimal visible data with motion blur and low light condition. (b) The blur removing process of transferring visible images into infrared style. The red box and blue box are visible content image and infrared style image, respectively. (c) The effect of texture details retaining by DSTFuse.

To tackle the shortcomings of conventional IVIF methods [16, 24, 36], numerous deep learning-based techniques have been developed. These method can be categorized into two main classes: Generative Adversarial (GANs)-based network [3, 32, 33, 46] and the Auto-Encoder-based network [13, 20, 21, 48]. The GAN-based methods typically consist of a generator responsible for generating the fused

image and a discriminator that evaluates the fusion performance. And Auto-Encoder-based methods extract features from infrared-visible images through an encoder and then map these features to a new representation space through a decoder. To leverage the the multimodal features, numerous previous works have attempted to map the most representative features of images from different modalities into the final image [11, 27, 40, 47, 55].

Despite a lot of researches on IVIF, there are few studies concentrating on utilizing suboptimal data, especially for data containing a significant number of blurry objects. The vast majority of studies utilize high-quality datasets such as TNO [43] and MSRS [41], which typically do not exhibit motion blur (Fig. 1(a)). However, the previous works using high-quality datasets have a limitation. Due to the variability of real-world environments, motion blur in source images is inevitable in practical applications of downstream tasks such as detection and segmentation. Therefore, it is crucial to mitigate the impact of a large volume of suboptimal data on IVIF tasks. Moreover, due to the significantly longer exposure time of RGB cameras compared to infrared cameras, the quality of infrared images for blurred objects in the same scene is superior to that of visible images. It is also a significant challenge to utilize the higher tolerance to blur that infrared images inherently possess due to differences in shooting equipment.

For the source visible and infrared images, the content information is intensely correlative. This is attributed to the high degree of coincidence in both the scene and the time of capture for each pair of infrared-visible images. It is intuitive that visible images, often prone to blurring due to equipment and target movement, have the potential to be transformed into consistently sharp infrared images. Previous studies on style transfer task [4, 14, 17, 45] have closely aligned with this concept.

In this paper, we present DSTFuse – a conceptually simple framework that aims to enhance deblurring via style transfer for IVIF. In DSTFuse, the blurry visible image is transformed into an image that combines infrared style with visible features by an Auto-Encoder-based cross-modality style transfer module (CST-module). Specifically, it aims to utilize infrared images as a reference to impose feature constraints on the blurry visible images, thus reducing motion-induced artifacts and enhancing details. Subsequently, DSTFuse utilizes the visible-infrared images to generate a fused image with rich background information and seamlessly integrates it with the output of CST-module into a meticulously crafted mapping function. As shown in Fig. 1(b), the contours of the blurred object in visible images under low-light conditions are gradually outlined, and details are filled in as the style transfer process. Moreover, the details in the fused image are also remarkably retained (Fig. 1(c)). This approach effectively harnesses the strong correlation between cross-modal images and the capability of style transfer to adapt to different modalities. The contributions of this work can be summarized in three aspects:

- We propose a dual-branch CNN-based framework for deblrring local blurry target and extracting and fusing global information, which better reflects the correspondence between modalities.
- We propose a style transfer module for the IVIF task to deblur the blurry target and retain visible information.
- Our method achieves promising image fusion results and also performs more superior in downstream tasks such as detection and segmentation.

2. Related Work

2.1. Infrared-visible fusion

With the development of deep learning, numerous work on IVIF task have emerged [11, 27, 40]. Ma *et al*. [33] proposed a GAN for IVIF task, conceptualizing the fusion algorithm as an adversarial game between retaining infrared thermal radiation information and maintaining visible appearance texture information, and achieved substantial breakthroughs. Then, Zhao *et al*. [55] pioneered the exploration of the two-scale decomposition in IVIF task, utilizing an encoder to decompose the images into background feature maps and detail feature maps, followed by a decoder used to reconstruct the original image. Recently, considering the combination of fusion and downstream pattern recognition tasks, Sun *et al*. [38] and Tang *et al*. [40] proposed the network driven by the downstream task and achieved promising results. Additionally, incorporating a pre-processing registration module before the fusion module has been shown to effectively address the misregistration of source images [11]. Zhao *et al*. [52] introduced a dual-branch Transformer-CNN network to correlate global and local features, achieving a fusion process where low-frequency features are related and high-frequency features are unrelated.

2.2. Style transfer

Style transfer, initially proposed by Leon *et al*. [4], aims to transfer the artistic style of one image onto another, creating an image with a unique artistic flair. Due to its innovative nature, this technique has attracted significant attention, then numerous style transfer models are implemented and utilized in various field [14, 15], particularly in image restoration and video processing. For image transformation problems, where an input image is converted into an output image, perceptual loss [17] has been designed and utilized for style transfer tasks. Then, Xun *et al*. [10] achieved arbitrary style transfer in real-time by introducing a novel adap-

tive instance normalization. To tackle the chanllenge of versatile style transfer, Wu *et al*. [45] implemented video style transfer without video in training process through InfoNCE loss [44]. Recently, Kwon *et al*. [18] proposed a network called CLIPstyler, capable of performing style transfer with just a single text condition, achieving results comparable to other models that use more complex inputs. The fundamental principle of classical style transfer methods is to generate an image that preserves the content of the original image while seamlessly incorporating the distinctive characteristics of the target style. This ensures that visible images retain more visually detailed texture during the deblurring process.

3. Method

The DSTFuse mainly consists of three modules, which are detailed in Fig. 2. In the cross-modality style transfer module (CST-module), the original visible image is combined with the edge information generated by the Sobel algorithm [6] as input. This concatenated input is then fed into the Auto-Encoder-like module to generate a structure-clear image that is similar in style to an infrared image. Finally, the infrared style image re-enters the CST-module as input to generate a new infrared-like image with more visible features. In the fusion module, the pipeline aims to train a Auto-Encoder-based structure for extracting features and reconstructing original images (in reconstruction stage) or generating fusion images initially (in fusion stage). In the mapping module, the output images from the fusion module and the CST-module are merged through an attention block to generate the final output image. The detailed workflow is illustrated in Fig. 2.

3.1. Cross-modality style transfer module

The CST-module aims to retain the visual effect of the visible image while minimizing motion blur as much as possible. To achieve this, the CST-module divides the training into two stages, focusing more on the infrared information in the first stage and the visible information in the second. In order to deblur efficiently, it adds feature constraints similar to style transfer to guide training of model and incorporated edge information $\mathcal{D}_S$ obtained from the Sobel algorithm $\mathcal{S}$ in both stages:

$$\mathcal{D}_S = \mathcal{S}(I) \oplus V, \tag{1}$$

where $\mathcal{S}(I)$ means the result of the Sobel algorithm [6] on the infrared image, which only retains the structure of the objects. $\oplus$ means element-wise addition.

Stage-1. Considering the focus of the first stage is the information of infrared image, the input of the first EBlock in encoder is designed as the concatenation of visible image V and edge information $\mathcal{S}$ to obtain more structural information. In addition, the edge information $\mathcal{S}$ is extracted as shallow features ϕ_S through a convolution. Then, the ϕ_S is used as the input, together with the second-to-last skip connection, into the final layer of the decoder.

Stage-2. After obtaining the image with more functional highlight and less motion blur, the output of the first stage serves as the input for the second stage. Different from the first stage, $\mathcal{S}$ is no longer used as an input so that the final output image will not contain highlighted edges.

The CST-module eliminate the motion blur of the target by introducing the style of infrared images. At the same time, the output image should not retain an excessive amount of visual information from the input image. Therefore, the perceptual loss [17] perfectly meets the requirements. The CST loss is:

$$L_{CST} = \alpha_1 L_{perceptual}(D, I, i) + \alpha_2 L_{SSIM}(F, V), \tag{2}$$

where $L_{perceptual}(D, I, i) = \|\phi(D, i) - \phi(I, i)\|_2$, and D is the output of CST-module, $\phi(\cdot, i)$ is the first i layers of a simple model extractor similar to VGG. As the i increases, the features become more abstract and the style becomes more biased towards the infrared image. In first stage of CST-module, it's need to retaining the structure of targets and reduce blurriness, while in the second stage, the focus is on retaining color, texture. Therefore, the layer of model extractor in the first stage is more than the second stage.

3.2. Fusion module

Reconstruction stage. The key to make Auto-Encoder perform better in image fusion is to extract the most representative features from source images. Capturing accurately feature that precisely reflects the advantage of visible and infrared images poses a significant challenge. And directly extract such features using a randomly initialized encoder instead of a well-pretrained one is not feasible.

To address this issue, the reconstruction stage is scheduled before the fusion stage. In this stage, a encoder is trained to extract features and a decoder to reconstruct them into original images for the subsequent fusion stage. Specifically, for the input image, it will pass through the encoder containing three EBlocks and the decoder with two DBlock and one OutBlock to reconstruct itself. The block struct can be seen in Fig. 2. And for each block, the residual-connection is used to accelerate convergence. In addition, skip connections between the first and last layers, and between the second and second-to-last layers, prevent gradient vanishing.

Since the aim of the reconstruction stage is to minimize the information loss of source image, the loss of reconstruction can be defined as:

$$L_{reconcstruct} = \alpha_1 f(I, \hat{I}) + \alpha_2 f(V, \hat{V}), \tag{3}$$

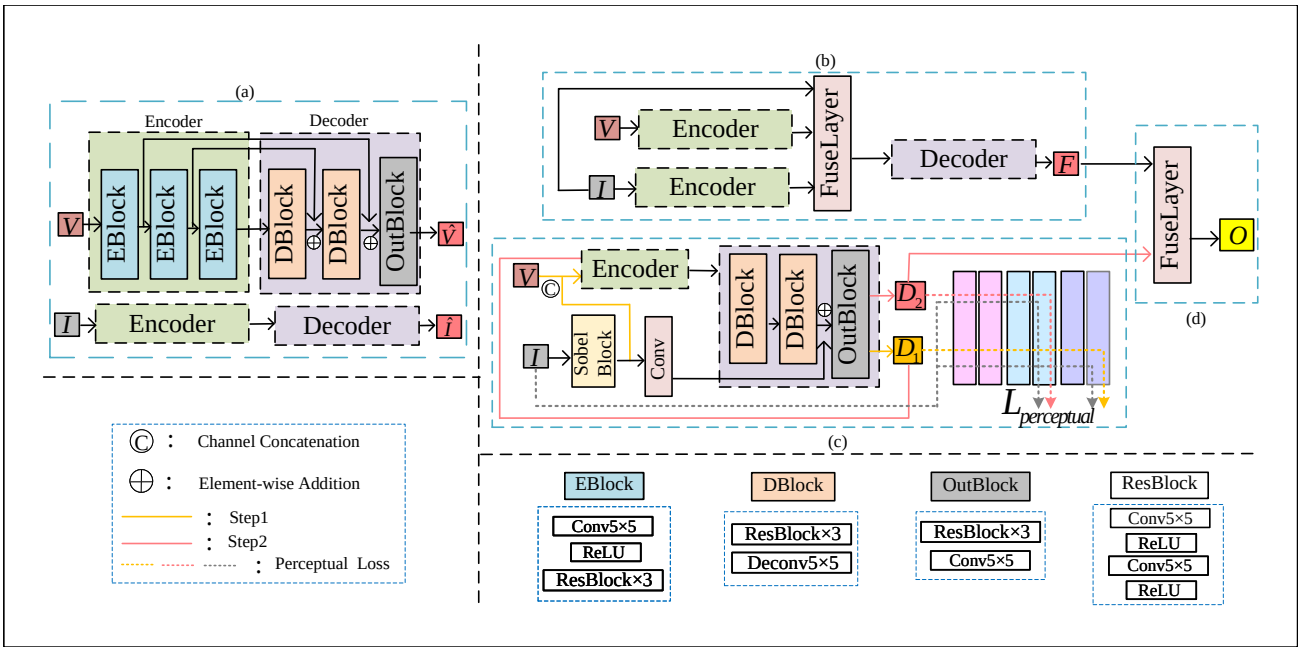

Figure 2. The architecture of DSTFuse, (a) The reconstruction stage of fusion module. (b) The fusion stage of fusion module. (c) The cross-modality style transfer module. (d) The mapping module.

where I and $\hat{I}$, V and $\hat{V}$ represent the input and output of infrared and visible images, respectively.

$$f(X,\hat{X}) = \|X - \hat{X}\|_2 + \lambda L_{SSIM}(X,\hat{X}), \tag{4}$$

where X and $\hat{X}$ represent the above input and output image, and $L_{SSIM}(X,\hat{X}) = 1 - SSIM(X,\hat{X})$. SSIM is the structural similarity index, which is a measure of the similarity between two pictures.

Fusion stage. After the reconstruction stage, the well-trained feature extractor $\mathcal{E}(\cdot)$ can be obtained. And the feature $\{\phi_V, \phi_I\}$ can be extracted from visible and infrared input $\{V, I\}$ by:

$$\phi_V = \mathcal{E}(V), \quad \phi_I = \mathcal{E}(I). \tag{5}$$

In previous studies, the neglect of suboptimal data has resulted in poor performance on datasets containing blurry images. In contrast to these work, the fusion module in DSTFuse is designed to prioritize the incorporation of detailed background information into the fused image, while deliberately disregarding the target object, which is instead the central focus of the CST module. Considering the high correlation between source visible and infrared images, it can be assumed that objects which appear motion-blurred in the visible image correspond to high-contrast and distinct targets in the infrared image. Therefore, the decoder should be prompted to learn the environmental information excluding the high-contrast targets. To this end, a fusion layer with attention block is added to highlight the background. And the mapping function is described as follow:

$$\phi_A = (\phi_V \oplus \phi_I) \oplus (\phi_V \oplus \phi_I) \otimes (1 - \mathcal{A}(I)), \tag{6}$$

where ϕ_V and ϕ_I are the features extracted from visible and infrared input, respectively. $\oplus$ and $\otimes$ means element-wise addition and element-wise multiplication. $\mathcal{A}(\cdot)$ is attention map matrix.

Finally, the output image F will preserve more detailed textures which is constrainted by the Sobel algorithm [6] and the gradient information. Additionally, the output should be similar to the visible image, so the loss function is:

$$\begin{aligned} L_{\text{fuse}} =& \alpha_1 \,\text{Sobel}(F,V,I) + \alpha_2 \| \,\text{F} - \max(V,I)\|_1 \\ &+ \alpha_3 L_{SSIM}(F,V), \end{aligned} \tag{7}$$

$$\text{Sobel}(F,V,I) = \text{Sobel}(F) - \max(\text{Sobel}(V), \text{Sobel}(I)), \tag{8}$$

where $\text{Sobel}(\cdot)$ is the Sobel algorithm [6].

In addition, the fusion module and the CST-module can be trained simultaneously.

3.3. Mapping module

After training through the fusion module and the CST-module, it is possible to obtain a fused image with detailed

environmental information and a small amount of functional highlights, as well as a infrared style image with a clear target structure and no motion blur. To integrate the benefits of both images into a final composite, the mapping module generates an attention map matrix derived from the infrared input. This matrix emphasizes the edges of all targets present in the scene. The mapping function is:

$$O = (D_2 \oplus V) \otimes \mathcal{A}(I) \oplus (F \oplus V) \otimes (1 - \mathcal{A}(I)), \quad (9)$$

where D_2 and F are the outputs of CST-module and fusion module, V and I is the input of visible and infrared image, respectively. $\oplus$ and $\otimes$ means element-wise addition and element-wise multiplication. $\mathcal{A}(\cdot)$ is the attention block.

After mapping, blurry parts of the final image are composed of the deblurred image, while the rest is composed of the fused image. The attention loss prompts the mapping matrix to focus only on the edges of the image, similar to fusion loss, which should be constrained by gradient and edge information:

$$L_{map} = \alpha_1 \, \mathrm{Sobel}(F, V, I) + \alpha_2 \| \, \mathrm{F} - \max(V, I) \|_1. \quad (10)$$

4. Experiment

4.1. Settings

Dataset and metrics. To verify the performance of model on deblurring, we select images with motion blur from the LLVIP dataset [12] as training set (317 pairs) and test set (60 pairs).

There are eight metrics used to quantitatively measure the fusion results: spatial frequency (SF), average gradient (AG), mean square error (MSE), peak signal to noise ratio (PSNR), mutual information (MI), visual information fidelity (VIF), correlation coefficient (CC), and structural similarity index measure (SSIM). The details of these metrics can be found in [30].

Implement details. DSTFuse is trained by Pytorch on single NVIDIA GeForce RTX 3090 GPU and Intel Xeon Gold 6330 CPU. The training samples are converted to grayscale images and resized to 640 × 640 in the preprocessing stage. In the training process, the Adam optimizer is employed, initializing the learning rate at 10^{-4}. The total number of training epochs is set to 15. During the first ten epochs, both the fusion module and the CST-module undergo concurrent training, with each of the reconstruction and fusion stages receiving training for precisely three epochs. In the final five epochs, the training is solely directed at the mapping module. For the tuning parameters in loss function, in Eq. (2), α_1 and α_2 are set to 100 and 1. In Eq. (3), α_1, α_2 and λ are set to 1, 1 and 5. In Eq. (7), α_1 to α_3 are set to 10, 5 and 1. In Eq. (10), α_1 and α_2 are set to 10 and 1.

4.2. Comparison with SOTA methods

In this section, DSTFuse is tested on the test set and compare the fusion results with the state-of-the-art methods including DIDFuse [55], RFN-Nest [22], MFEIF [27], ReCoNet [11], SeAFusion [40], DeFusion [25], MetaFusion [51], LLRNet [23] and EMMA [53].

Qualitative comparison. It has been shown the qualitative comparison in Fig. 3. Obviously, the proposed method more effectively integrates thermal radiation information from infrared images with detailed textures from visible images. As show in visual comparison result, the background information that was easily overlooked in previous methods due to the prominence of infrared images is perfectly retained in DSTFuse. This can be attributed to the CST-module, which does not forcibly merge visible images with infrared image , but rather performs only style conversion, thereby preserving most of the visible details. Consequently, for the objects in dark regions, DSTFuse appropriately highlights them for identification in downstream task. For blurry object, DSTFuse providing details that conform to human visual perception.

Quantitative comparison. Afterward, we follow the previous IVIF works by reporting eight metrics for visual evaluation criterion. There are excellent performance across most metrics, demonstrating that it is suitable for the human visual perception without bias from observers or interpreters. Specifically, the optimal results on MI and CC [30] show that the fused image contain the most amount of information and the strongest correlation between source images and fused image, respectively. Besides, the promising result on SF, AG, MSE, PSNR and VIF [30] indicates show that the proposed fusion method produces the most texture details, least distortion and best matches to the human visual system.

Visualization of CST-module. Fig. 4 visualizes the effectiveness of perceptual loss in CST-module. Obviously, with training goes on, more detail texture of target are activated and more background information are inactivate. As the input of CST-module, the visible image contains the abundant details of the target and exhibit significant perceptual differences compared to the infrared images which is regarded as style image in style transfer. In the group of CST output, the CST-module firstly focus on the profile of target, showing that the deblurring function works well. As the perceptual loss reaches convergence, an increasing amount of detail is incorporated.

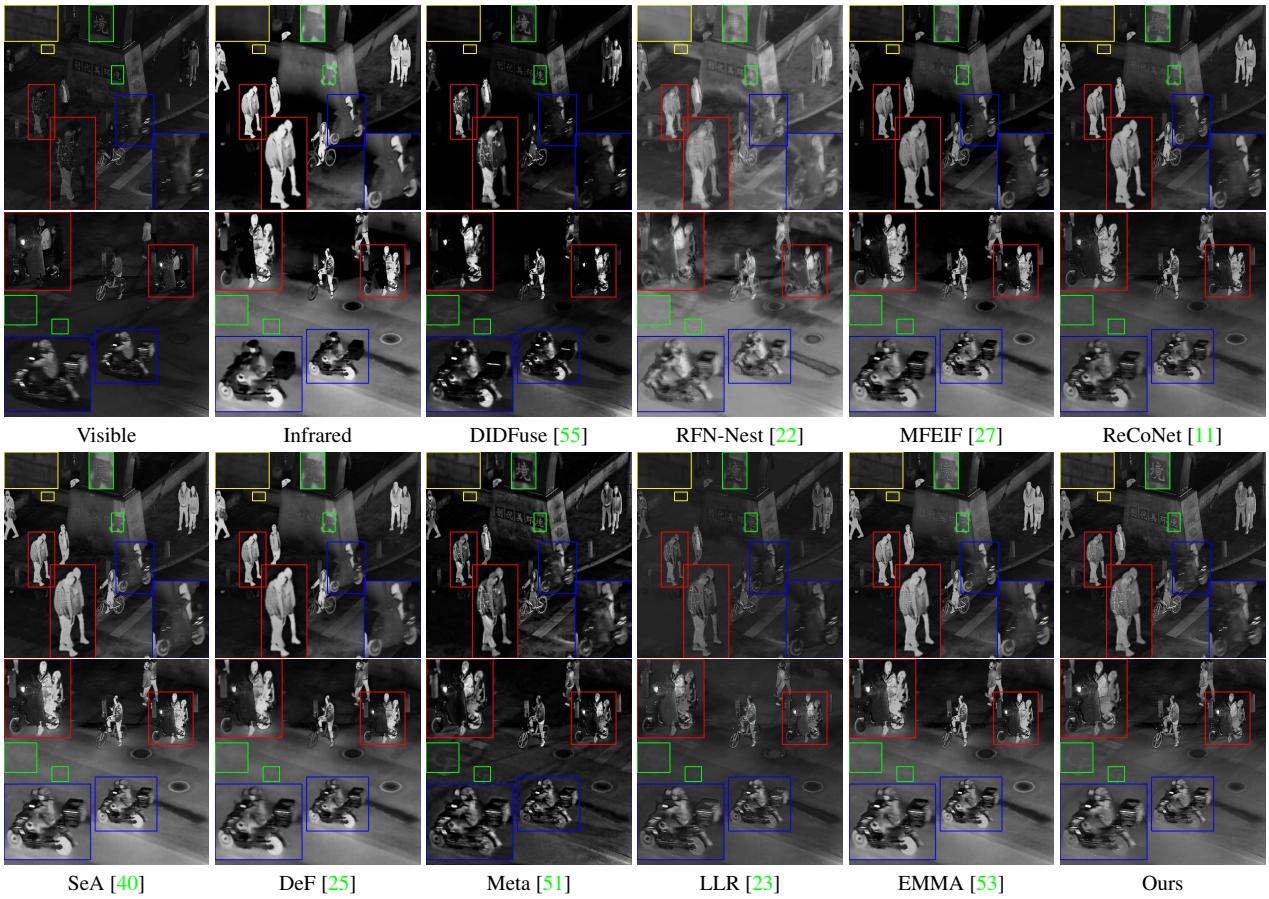

Figure 3. Visual comparison for "010018" (up) and "050131" (down) in LLVIP IVIF dataset.

	DID [55]	RFN [22]	MFE [27]	ReC [11]	SeA [40]	DeF [25]	Meta [51]	LLR [23]	EMM [53]	Ours
SF	4.943	4.818	4.567	4.712	5.440	4.808	**6.137**	4.528	5.469	<u>5.582</u>
AG	2.036	2.251	1.882	2.031	2.432	2.148	**2.985**	1.787	2.430	<u>2.540</u>
MSE	0.043	0.060	**0027**	0.030	0.037	0.036	0.037	0.031	0.030	<u>0.029</u>
PSNR	13.71	12.22	**15.78**	14.26	14.59	14.40	14.46	14.83	15.42	<u>15.56</u>
MI	1.581	<u>1.601</u>	1.181	1.357	1.506	1.286	1.214	1.494	1.595	**1.650**
VIF	0.746	0.753	0.814	0.813	**0.934**	0.850	0.898	0.593	0.903	<u>0.919</u>
CC	0.686	0.674	<u>0.712</u>	0.707	0.696	0.647	0.684	0.693	0.703	**0.734**
SSIM	1.044	0.999	1.298	**1.398**	1.358	<u>1.367</u>	1.206	1.304	1.306	1.316

Table 1. Quantitative results of the IVIF task. The **Bold** and <u>underline</u> show the best, second-best value, respectively.

4.3. Ablation studies

The ablation studies are conducted on the LLVIP dataset [12] to prove the rationality of DSTFuse, with the results shown in Tab. 2 and Fig. 5.

Essential module in DSTFuse. To independently validate the efficacy of the fusion module and the mapping module, two comparative experiments have been devised. In Exp. I, the mapping module is removed to ascertain its capability in accurately mapping cross-modal information. As an alternative, the summation method is used to integrate output of CST-modlue with that of the fusion module. In formula, the summation method can be described as:

$$O = (D_2 \oplus F), \tag{11}$$

where D_2 and F are the outputs of CST-module and fusion module, respectively.

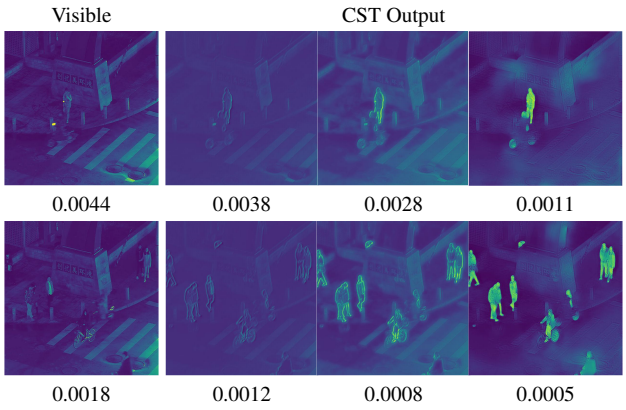

Figure 4. Visualization of the CST-module for "010018" (up), "010054" (down) in LLVIP IVIF dataset. The values represent the results of the perceptual loss.

When removing the mapping module, although the network retains the ability to execute feature mapping, it falls short in precisely selecting the requisite information from distinct images. In Exp. II, the fusion module is eliminated to confirm the proficiency in extracting background information. To substitute for the output of this module, the original visible image is utilized to provide background information. Results in Exp. II illustrate that the absence of effective feature extraction results in a lack of detail and texture, particularly in darker regions, thereby causing a decline in overall performance.

Term in loss function. Then, in Exp. III, it separately removes the perceptual loss from CST-module and modifies it to adopt the conventional loss function used in other fusion tasks, denoted as $\mathcal{L}_2 = \|x - \hat{x}\|_2$. And in the first step of CST-module, the x represents the infrared image, while in the second step, it represents the visible image. The perceptual loss ensures that, during the style transfer process within the CST-module, the content image adequately inherits information from the style image, thereby making the generated image perceptually more similar to the source image. In contrast, the conventional loss function merely enforces the image to be similar to the source image. Results in Exp. III demonstrate the necessity of perceptual loss.

4.4. Application in the downstream taks

To evaluate the promoting effect of fused image and its improved performances on downstream task, further external validation is conducted. For infrared-visible object detection, the fused images generated by state-of-the-art models are evaluated using five classic detectors by comparing the AP value for person detection. The selected detectors include Faster R-CNN [5], YOLOv5 [37], SSD [28], RetinaNet [26] and Mask R-CNN [7]. For infrared-visible semantic segmentation, the segmentation network includes FCN [29], DeeplabV3 [2] and LSR-APP [8]. And the performance is evaluated using Intersection over Union (IoU) for person segmentation.

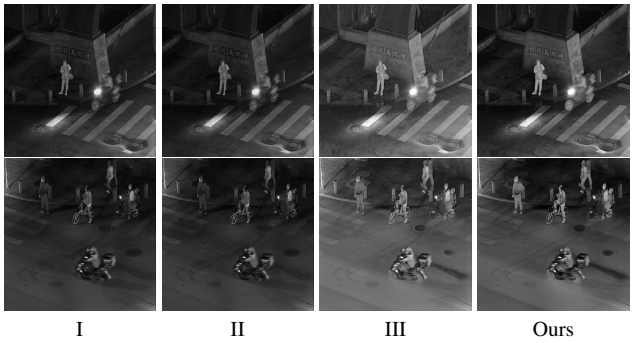

Figure 5. Ablation experiment for "010562" (up) and "050131" (down) in LLVIP IVIF dataset.

	Configurations	SF	PSNR	CC	VIF
I	w/o Mapping Module	5.364	14.36	0.637	0.645
II	w/o Fusion Module	4.947	14.24	0.680	0.835
III	w/o Perceptual Loss	5.474	14.15	0.730	0.835
	Ours	**5.582**	**15.56**	**0.734**	**0.919**

Table 2. Ablation experiment results. **Bold** indicates the best value.

Object detection. As shown in Tab. 3, DSTFuse plays a significantly positive role in detection. In comparison to direct predictions made on the source images, the fused images generated by DSTFuse substantially enhance prediction accuracy across all five detection models. Compared to previous work, DSTFuse exhibits the promising superior detection capabilities, which can be contributed to its ability to preserve information that aligns closely with the human visual system. To obtain a more intuitive comparison, the detection results are compared using YOLOv5 [37] as the detector and the visual results are shown in Fig. 6. In the first example, when the infrared source images have already been effectively detected, only the fused image generated by DSTFuse can retain the infrared features and be detected. And for those infrared images that perform poorly in detection due to overly prominent functional highlights (*e.g.*, the second example in Fig. 6), the fused images generated by DSTFuse appropriately balance the high contrast of the targets with the real pixel intensity. This allows the detector to accurately identify each target.

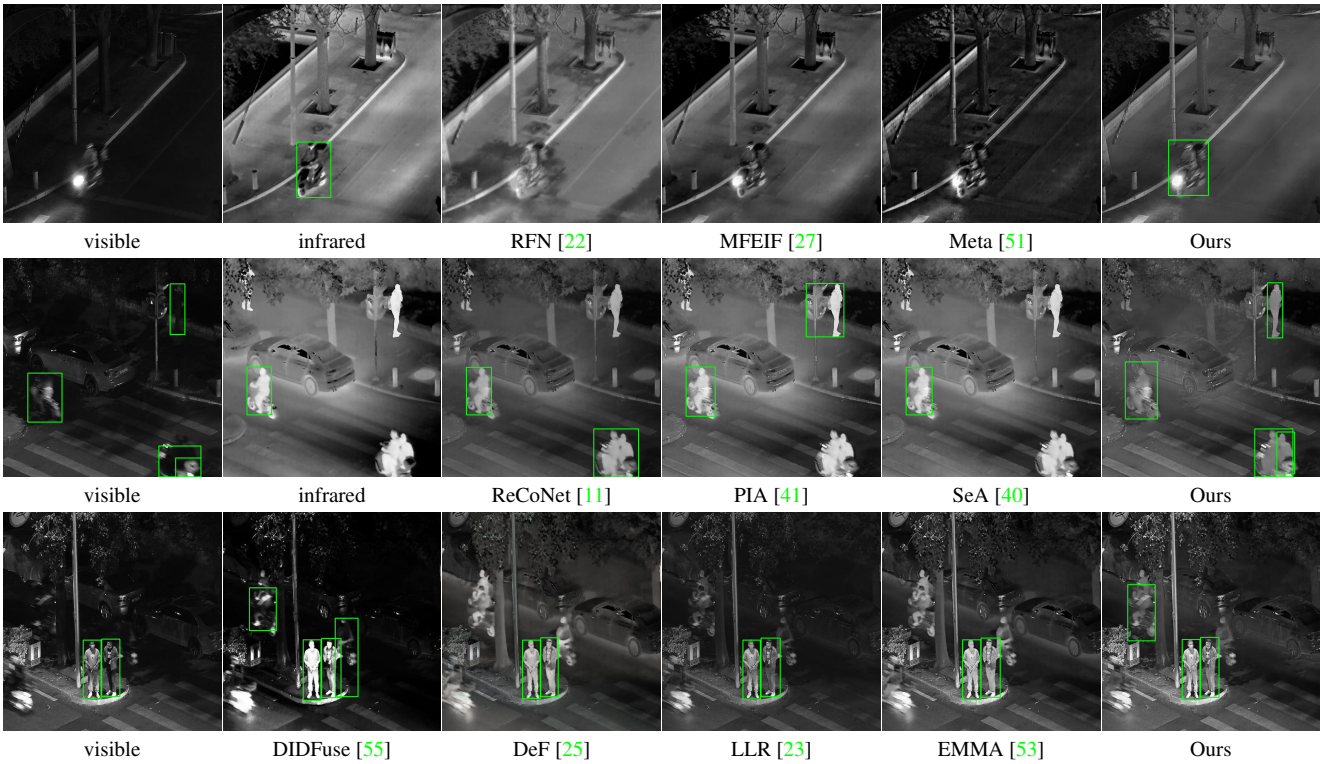

Figure 6. Detection results on source images and different fused image for "040145" (up), "080246" (middle) and "080786" (down) in LLVIP IVIF dataset.

	Faster RCNN [5]	YOLOv5 [37]	SSD [28]	Retina [26]	Mask RCNN [7]
VI	30.52	39.26	40.40	43.08	55.20
IR	34.03	37.80	39.29	41.28	48.97
DID [55]	32.05	39.11	44.97	43.56	54.38
RFN [22]	15.15	27.65	16.52	27.63	33.88
MFE [27]	32.48	42.57	39.67	41.00	50.94
ReC [11]	**37.59**	46.10	46.65	42.23	54.68
SeA [40]	36.25	44.54	46.13	46.84	55.24
DeF [25]	37.57	44.31	43.45	46.60	52.82
Meta [51]	36.43	43.49	**49.36**	**52.19**	56.42
LLR [23]	33.55	**53.37**	45.02	45.90	52.83
EMMA [53]	34.34	45.91	45.02	43.67	53.16
Ours	36.58	46.87	42.85	48.86	**57.24**

Table 3. AP(%) values of person for detection on LLVIP dataset. The **bold** and underline show the best and second-best value, respectively.

	FCN [29]	DeeplabV3 [2]	LSR-APP [8]
DID [55]	47.91	48.12	48.07
RFN [22]	44.69	46.49	47.99
MFE [27]	48.23	48.51	48.37
ReC [11]	48.32	**48.70**	**48.57**
SeA [40]	48.34	48.63	48.53
DeF [25]	48.37	48.52	48.52
Meta [51]	47.99	48.32	48.26
LLR [23]	47.97	48.29	48.13
EMMA [53]	48.10	48.37	48.41
Ours	**48.48**	48.64	48.48

Table 4. IoU(%) values of person for semantic segmentation on LLVIP dataset. The **bold** and underline show the best and second-best value, respectively.

Semantic segmentation. To evaluate the performance of DSTFuse on infrared-visible semantic segmentation, we selected 42 pairs of infrared and visible images from the LLVIP dataset [12], and proceeded to annotate the person category within these images. The result in Tab. 4 show that DSTFuse effectively integrates the contour details from the source images, thereby enhancing the model's ability to recognize object boundaries and achieving more accurate segmentation.

5. Conclusion

This paper presents a infrared-visible fusion framework through introducing the style transfer. With the cross-modality style transfer module, target with motion blur in

visible image are more clearly outlined and more easily recognized. Experiments demonstrate the fusion effect of DSTFuse, and the performance on downstream detection and segmentation can be also improved.

References

[1] Njuod Alsudays, Jing Wu, Yu-Kun Lai, and Ze Ji. Afpsnet: Multi-class part parsing based on scaled attention and feature fusion. In *WACV*, pages 4033–4042, 2023. 1

[2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 7, 8

[3] Yu Fu, Xiao-Jun Wu, and Tariq Durrani. Image fusion based on generative adversarial network consistent with perception. *Information Fusion*, 72:110–125, 2021. 1

[4] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, June 2016. 2

[5] Ross Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015. 7, 8

[6] Rafael C Gonzalez, Richard E Woods, and Steven L Eddins. *Digital Image Processing Using MATLAB*. Prentice Hall, 2009. 3, 4

[7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 7, 8

[8] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. In *ICCV*, October 2019. 7, 8

[9] Hai-Miao Hu, Jiawei Wu, Bo Li, Qiang Guo, and Jin Zheng. An adaptive fusion algorithm for visible and infrared videos based on entropy and the cumulative distribution of gray levels. *IEEE T MULTIMEDIA*, 19(12):2706–2719, 2017. 1

[10] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 2

[11] Zhanbo Huang, Jinyuan Liu, Xin Fan, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Reconet: Recurrent correction network for fast and efficient multi-modality image fusion. In *ECCV*, pages 539–555. Springer, 2022. 2, 5, 6, 8

[12] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *ICCV*, pages 3496–3504, 2021. 5, 6, 8

[13] Lihua Jian, Xiaomin Yang, Zheng Liu, Gwanggil Jeon, Mingliang Gao, and David Chisholm. Sedrfuse: A symmetric encoder–decoder with residual block network for infrared and visible image fusion. *IEEE T INSTRUM MEAS*, 70:1–15, 2020. 1

[14] Yongcheng Jing, Xiao Liu, Yukang Ding, Xinchao Wang, Errui Ding, Mingli Song, and Shilei Wen. Dynamic instance normalization for arbitrary style transfer. In *AAAI*, 2020. 2

[15] Yongcheng Jing, Yang Liu, Yezhou Yang, Zunlei Feng, Yizhou Yu, Dacheng Tao, and Mingli Song. Stroke controllable fast style transfer with adaptive receptive fields. In *ECCV*, 2018. 2

[16] Jing jing Zong and Tian shuang Qiu. Medical image fusion based on sparse representation of classified image patches. *Biomedical Signal Processing and Control*, 34:195–205, 2017. 1

[17] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711. Springer, 2016. 2, 3

[18] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In *CVPR*, pages 18062–18071, 2022. 3

[19] Fayez Lahoud and Sabine Susstrunk. Ar in vr: Simulating infrared augmented vision. In *ICIP*, pages 3893–3897. IEEE, 2018. 1

[20] Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE TIP*, 28(5):2614–2623, May 2019. 1

[21] Hui Li, Xiao-Jun Wu, and Tariq Durrani. NestFuse: An Infrared and Visible Image Fusion Architecture based on Nest Connection and Spatial/Channel Attention Models. *IEEE T INSTRUM MEAS*, 69(12):9645–9656, 2020. 1

[22] 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, March 2021. 5, 6, 8

[23] Hui Li, Tianyang Xu, Xiao-Jun Wu, Jiwen Lu, and Josef Kittler. LRRNet: A novel representation learning guided fusion framework for infrared and visible images. *IEEE TPAMI*, 45(9):11040–11052, 2023. 5, 6, 8

[24] Shutao Li, Bin Yang, and Jianwen Hu. Performance comparison of different multi-resolution transforms for image fusion. *Information Fusion*, 12(2):74–84, 2011. 1

[25] Pengwei Liang, Junjun Jiang, Xianming Liu, and Jiayi Ma. Fusion from decomposition: A self-supervised decomposition approach for image fusion. In *ECCV*, 2022. 5, 6, 8

[26] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 7, 8

[27] 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. *TCSVT*, 32(1):105–119, 2021. 2, 5, 6, 8

[28] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37. Springer, 2016. 7, 8

[29] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 7, 8

[30] Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: A survey. *Information Fusion*, 45:153–178, 2019. 5

[31] Jiayi Ma, Linfeng Tang, Fan Fan, Jun Huang, Xiaoguang Mei, and Yong Ma. Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer. *JAS*, 9(7):1200–1217, 2022. 1

[32] Jiayi Ma, Han Xu, Junjun Jiang, Xiaoguang Mei, and Xiao-Ping Zhang. Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE TIP*, 29:4980–4995, 2020. 1

[33] 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. 1, 2

[34] Bikash Meher, Sanjay Agrawal, Rutuparna Panda, and Ajith Abraham. A survey on region based image fusion methods. *Information Fusion*, 48:119–132, 2019. 1

[35] Lukas Mehl, Azin Jahedi, Jenny Schmalfuss, and Andrés Bruhn. M-fuse: Multi-frame fusion for scene flow estimation. In *WACV*, pages 2020–2029, 2023. 1

[36] Ujwala Patil and Uma Mudengudi. Image fusion using hierarchical pca. In *ICIIP*, pages 1–6. IEEE, 2011. 1

[37] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016. 7, 8

[38] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Detfusion: A detection-driven infrared and visible image fusion network. In *ACM MM*, pages 4003–4011, 2022. 2

[39] Linfeng Tang, Yuxin Deng, Yong Ma, Jun Huang, and Jiayi Ma. Superfusion: A versatile image registration and fusion network with semantic awareness. *JAS*, 9(12):2121–2137, 2022. 1

[40] Linfeng Tang, Jiteng Yuan, and Jiayi Ma. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Information Fusion*, 82:28–42, 2022. 1, 2, 5, 6, 8

[41] Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83-84:79–92, 2022. 2, 8

[42] Linfeng Tang, Hao Zhang, Han Xu, and Jiayi Ma. Deep learning-based image fusion: A survey. *Journal of Image and Graphics*, 28(1):3–36, 2023. 1

[43] Alexander Toet and Maarten A. Hogervorst. Progress in color night vision. *Optical Engineering*, 51:010901 – 010901, 2012. 2

[44] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. 3

[45] Zijie Wu, Zhen Zhu, Junping Du, and Xiang Bai. Ccpl: Contrastive coherence preserving loss for versatile style transfer. In *ECCV*, pages 189–206. Springer, 2022. 2, 3

[46] Han Xu, Pengwei Liang, Wei Yu, Junjun Jiang, and Jiayi Ma. Learning a generative model for fusing infrared and visible images via conditional generative adversarial network with dual discriminators. In *IJCAI*, pages 3954–3960, 2019. 1

[47] Han Xu, Xinya Wang, and Jiayi Ma. Drf: Disentangled representation for visible and infrared image fusion. *IEEE T INSTRUM MEAS*, 70:1–13, 2021. 2

[48] Han Xu, Hao Zhang, and Jiayi Ma. Classification saliency-based rule for visible and infrared image fusion. *IEEE TCI*, 7:824–836, 2021. 1

[49] Mingde Yao, Zhiwei Xiong, Lizhi Wang, Dong Liu, and Xuejin Chen. Spectral-depth imaging with deep learning based reconstruction. *Optics express*, 27(26):38312–38325, 2019. 1

[50] Hao Zhang, Han Xu, Xin Tian, Junjun Jiang, and Jiayi Ma. Image fusion meets deep learning: A survey and perspective. *Information Fusion*, 76:323–336, 2021. 1

[51] Wenda Zhao, Shigeng Xie, Fan Zhao, You He, and Huchuan Lu. Metafusion: Infrared and visible image fusion via meta-feature embedding from object detection. In *CVPR*, pages 13955–13965, 2023. 5, 6, 8

[52] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *CVPR*, pages 5906–5916, June 2023. 2

[53] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Kai Zhang, Shuang Xu, Dongdong Chen, Radu Timofte, and Luc Van Gool. Equivariant multi-modality image fusion. In *CVPR*, June 2024. 5, 6, 8

[54] Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, and Jiangshe Zhang. Bayesian fusion for infrared and visible images. *Signal Processing*, 177:107734, Dec. 2020. 1

[55] Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, Jiangshe Zhang, and Pengfei Li. Didfuse: Deep image decomposition for infrared and visible image fusion. In *PRICAI*, IJCAI-PRICAI-2020. IJCAI, July 2020. 2, 5, 6, 8