DRF: Disentangled Representation for Visible and Infrared Image Fusion

Han Xu, Xinya Wang, and Jiayi Ma, *Member, IEEE*

***Abstract*— In this article, we propose a novel decomposition method by applying disentangled representation for visible and infrared image fusion (DRF). According to the imaging principle, we perform the decomposition depending on the source of information in the visible and infrared images. More concretely, we disentangle the images into the scene- and sensor modality (attribute)-related representations through the corresponding encoders, respectively. In this way, the unique information defined by the attribute-related representation is closer to the information captured by each type of sensor individually. Thus, the problem of inappropriate extraction of unique information can be alleviated. Then, different strategies are applied for the fusion of these different types of representations. Finally, the fused representations are fed into the pretrained generator to generate the fusion result. The qualitative and quantitative experiments on the publicly available *TNO* and *RoadScene* data sets demonstrate the comparable performance of our DRF over the state of the art in terms of both visual effect and quantitative metrics.**

***Index Terms*— Deep learning, disentangled representation, image fusion, infrared, visible.**

I. Introduction

IMAGE fusion is an active research topic with a wide variety of applications, including security, industrial, civilian, and medical fields [1], [2]. In the image fusion problem, because of the limitation of hardware devices, different sensors can merely capture a part of the scene information. Specifically, in the visible (VIS) and infrared (IR) image fusion (VIF) problem, the visible sensor captures the reflected light information but is greatly affected by lighting and occlusion. By contrast, the infrared sensor captures the thermal radiation information, but the infrared image usually has the disadvantages of serious noise and few details. As the source images can complement each other, the purpose of image fusion is to extract the vital and complementary information/features from the source images and use them to generate a single fused image. Then, the fused image with more comprehensive scene information, superior visual perception, and higher target saliency is more suitable for subsequent processing or visual tasks, e.g., target recognition, classification, and object detection [3], [4]. Thus, VIF can play a role in surveillance, vehicle navigation, and monitoring by breaking through the barrier of illumination and occlusion. A diverse range of applications is benefiting from the fusion operation [5].

In fact, the infrared images obtained from thermal infrared instruments always suffer from serious noise and blurred details due to the imperfection of instruments. To improve the imaging quality of instruments, we need to overcome the objective impact of impulse noise or bear high costs, which is challenging or expensive. By comparison, the development of visible imaging instruments is comparatively more perfect. They have better imaging quality and lower cost. Thus, VIF can be regarded as using the visible instruments to enhance the thermal infrared instruments and improve their imaging quality.

To achieve the target of VIF, the VIF algorithms are devoted to the feature extraction from different source images and their fusion rules. Among them, many algorithms aim to decompose the source images into different parts to extract various features. Then, multiple fusion strategies are designed according to the characteristics of these parts. In these methods, in order to facilitate the manual design of fusion strategies, the extracted features are often of the same type or have the same meaning. According to the theory, these methods mainly include multiscale transform-based methods [6]–[8], sparse representation-based methods [9]–[12], and low-rank representation-based methods [13]–[15]. For instance, in the multiscale transform, the pyramid transform aims to decompose the source images into multiscales of spatial frequency bands, and the wavelet transform decomposes the source images into a series of high- and low-frequency subimages. The settings of the frequency bands are the same for different types of source images. In the sparse representation-based methods, different types of source images are sparsely represented by the same learned overcomplete dictionary and their respective sparse representation coefficients. In the low-rank representation, the low-rank structure and the salient component are decomposed from the source images. For extracting the salient component, a projection matrix named a salient coefficient matrix is learned and shared for different source images. However, even though the source images are decomposed into a series of parts, these methods still use the same representations in these decomposed components for VIS and IR images, regardless of their distinct modalities. For example, even though the wavelet transform decomposes the

Manuscript received November 27, 2020; revised January 5, 2021; accepted January 24, 2021. Date of publication February 3, 2021; date of current version February 19, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61773295 and in part by the Natural Science Foundation of Hubei Province under Grant 2019CFA037. The Associate Editor coordinating the review process was Yan Zhuang. *(Han Xu and Xinya Wang contributed equally to this work.) (Corresponding author: Jiayi Ma.)*

The authors are with the Electronic Information School, Wuhan University, Wuhan 430072, China (e-mail: xu_han@whu.edu.cn; wangxinya@whu.edu.cn; jyma2010@gmail.com).

Digital Object Identifier 10.1109/TIM.2021.3056645

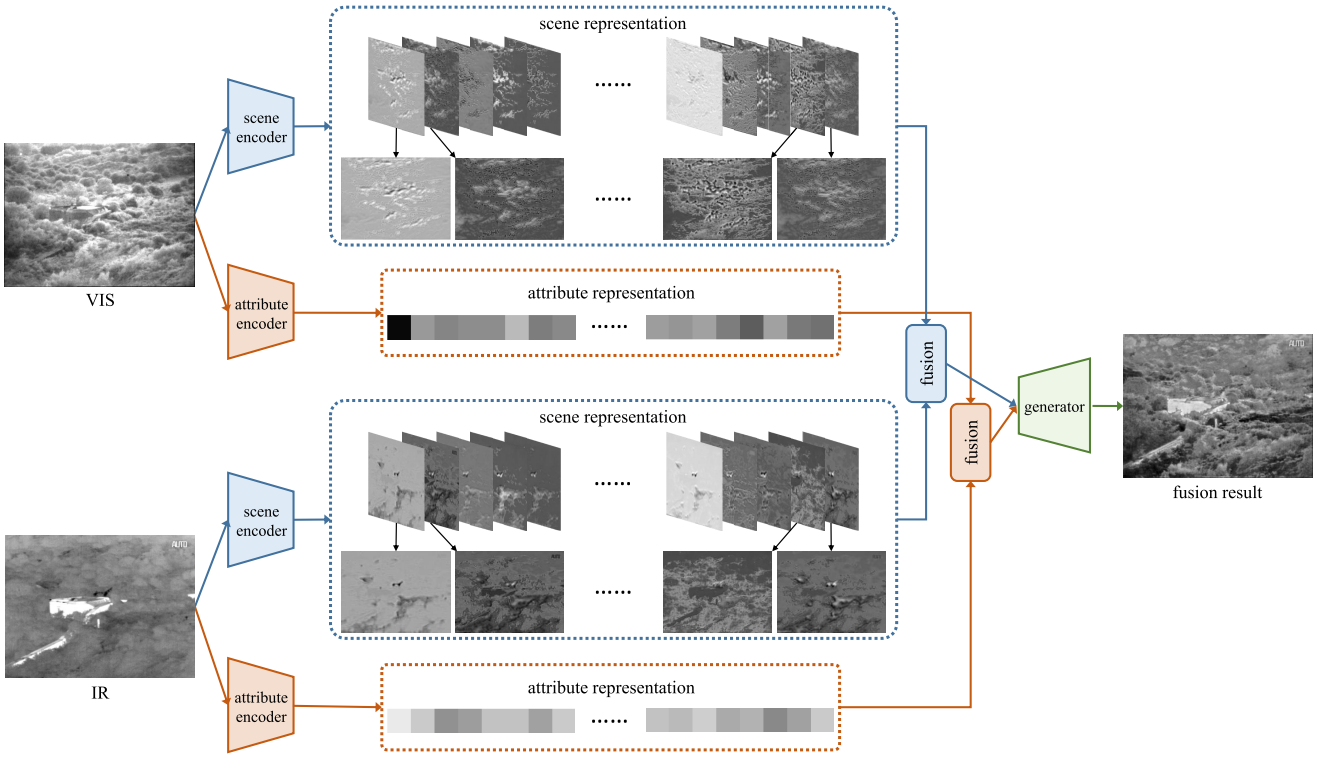

Fig. 1. Framework of the proposed DRF. The two scene encoders are pseudo-Siamese networks (they have the same network architecture but do not share the same weights) and so are the two attribute encoders. The attribute representation is a z-dimensional vector. In this figure, it is shown in four rows for ease of observation.

source images into different-frequency subimages, the subimages of the VIS and IR images in the same frequency are still of the same representation. Nevertheless, even if they are both high-frequency information, their physical meanings are vastly different. In the IR image, the high-frequency information represents the border of different materials or objects. For example, as shown in Fig. 1, the border of the bunker and the background is of high frequency, while the high-frequency information in the VIS image represents the abundant texture feature. Thus, both the high-frequency information in VIS and IR images should be retained in the fused image. However, the fusion process performed on this scale is bound to cause the distortion of some valuable information, while, in other scales, there may exist the situation where both the subimages of IR and VIS images contain little information. However, the existence of this band leads to the retention of less valuable information. Therefore, it is inappropriate to use the same representation for VIS and IR images as it may result in redundancy or distortion of information.

To alleviate the inappropriateness in the abovementioned algorithms, some methods apply different representations for VIS and IR images, respectively. They describe or split the unique information in each source image in manual ways. Among them, a series of methods use the pixel intensity distribution to describe the thermal radiation information in the IR image and characterize the reflected lighting information in the VIS image with gradients [16]–[20]. However, the manually designed splitting ways cannot completely characterize the unique information in each source image. For example, the gradients of the IR image also contain unique thermal radiation information.

To solve the problem, we aim to split the unique information from the common information in the source images as much as possible. For this purpose, we turn the attention back to the imaging process of source images. Whether the source images are captured from the visible or infrared sensor, they are shot from the same scene, which contains massive information. The difference is that these two types of sensors use their specific imaging modalities to capture a part of the original information. These imaging modalities can be regarded as the post-processing of the original massive information. Under the joint action of the captured scene and the specific imaging modality of the sensor, the VIS and IR images present the same scene with distinctive representations, including gradients, contrast, and illuminance. Therefore, instead of splitting the source images according to the forms of information representation, such as frequency, sparse coefficients, and salient components, we perform the decomposition according to the source of the information. More concretely, we decompose a source image into two parts: the information from the scene and that related to the sensor modality. As the information related to the sensor modality reflects the attributes of the sensors or the source images, we define this type of information as the unique attribute representation, while the information from the scene, i.e., the scene representation, is the common information of both types of source images. Based on this novel decomposition way, we propose a new VIF method, termed *disentangled representation for visible and infrared image fusion (DRF)*.

In DRF, we apply the disentangled representation to disentangle the scene and attribute representations in the source images. The scene representation is extracted as the common information with the scene encoder, and the attribute representation is extracted as the unique information with the attribute encoder, as shown in Fig. 1. In this way, the unique information can be split from the common information with interpretable physical meaning as the unique information is related to the imaging modality. Then, appropriate fusion strategies are performed on the disentangled scene and attribute representations of different source images separately. Finally, the fusion result can be generated through a pretrained generator according to the fused representations. The qualitative and quantitative results show that our method can achieve comparable performance. The contributions of our work include the following two aspects.

1) We introduce a novel decomposition method for image fusion. We propose a new viewpoint that the source images are formed by the joint action of scene and sensor modality. Based on it, we decompose the source images according to the sources of the information rather than the forms of information representation in existing decomposition-based fusion methods.
2) From the abovementioned viewpoint, we introduce the disentangled representation for image fusion. We disentangle the VIS and IR images into the scene- and attribute-related representations through the encoders. Then, different strategies are applied for the fusion of these representations, respectively. Finally, the fused representations are fed into a pretrained generator to generate the fusion result. Thus, each network in our method also has better interpretability.

The remainder of this article is arranged as follows. Section II provides some discussions about related work, including infrared and visible image fusion methods and disentangled representation-related methods. In Section III, we give a detailed introduction to our proposed DRF with the problem formulation, loss functions, architectures, and other implementations in detail. In Section IV, the qualitative and quantitative evaluations and the experiment of the hyperparameters are performed. Section V gives the conclusion.

II. Related Work

A. Visible and Infrared Image Fusion Methods

Besides the abovementioned fusion methods in Section I, there are also some methods based on other theories. For instance, Han *et al.* [21] proposed a saliency-aware fusion algorithm to enhance the visualization of the VIS image with the object information in the IR image. It first applies a saliency detection to depict the foreground object in the IR image. Then, it biases the final result in favor of the VIS image, except the region with clear thermal saliency. On the contrary, GTF [16] aims to enhance the IR image with the abundant textures in the VIS image. Thus, the superior contrast of the IR image is preserved to highlight the thermal targets. More specifically, it characterizes the reflected light information in VIS images with gradients and uses the pixel intensity distribution to describe the thermal radiation information/contrast in IR images.

With the significant progress of deep learning in computer vision, scholars have proposed many image fusion methods that are based on deep learning. For instance, Li *et al*. [22] decomposed the source images into detail and base contents. A deep learning framework is employed to extract multiple features from the detailed content. Based on them, it generates some candidates for the final detail content and selects as the final one through the max selection strategy. In [23], a method was proposed to generate fusion results with high similarity to RGB images. At the same time, pedestrian visibility is enhanced by training a network to learn some relevant features of human appearance. Zhang *et al.* [24] tried to fuse RGB and IR images. It first extracts coarse features from RGB/IR images and further extracts multilevel refined features. Then, these features are fused at each level to generate the cross-modal features through a multibranch module. Finally, multilevel fused features are integrated. In these methods, the networks are employed to extract deep-level or multilevel features from original source images without considering their meanings.

Also, according to the theory of generative adversarial networks (GANs), the researchers have proposed several GAN-based VIF methods. For example, FusionGAN [18] uses a generator to generate the fused image. Meanwhile, a discriminator is trained to distinguish differences between the result and the visible image. In this way, the fused image is forced to have more details of the visible image. Based on it, DDcGAN [20] adds an infrared discriminator to distinguish differences between the result and infrared image and improves the generator architecture to solve the multiresolution image fusion issue. Besides, ResNetFusion [19] designed two additional loss functions. A detailed loss is used to improve the detail quality, and a target edge-enhancement loss helps sharpen the edges of thermal targets. Moreover, AttentionFGAN [25] integrated multiscale attention mechanism. It helps the generator focus on the target in the foreground and background details. The discriminators focus on attention regions. However, these manually designed splitting ways (such as pixel intensity distribution and gradients) cannot completely characterize the unique information in each source image.

B. Disentangled Representation

Disentangled representation is a theory that aims to model the factors of data variations. Then, a disentangled representation of the input image can be learned. Some works have tried to apply the disentangled representation in the computer vision community. Tran *et al*. [26] proposed DR-GAN to learn the explicit disentangled representation from face variations. Based on this representation, the pose-invariant face recognition can be realized. Lee *et al.* [27], [28] explicitly embedded images into two spaces: domain-invariant content space and domain-specific attribute space. By changing the domain-specific attributes, it can realize the image translation between two visual domains. For single-image deblurring, Lu *et al*. [29] disentangled the content features and blur

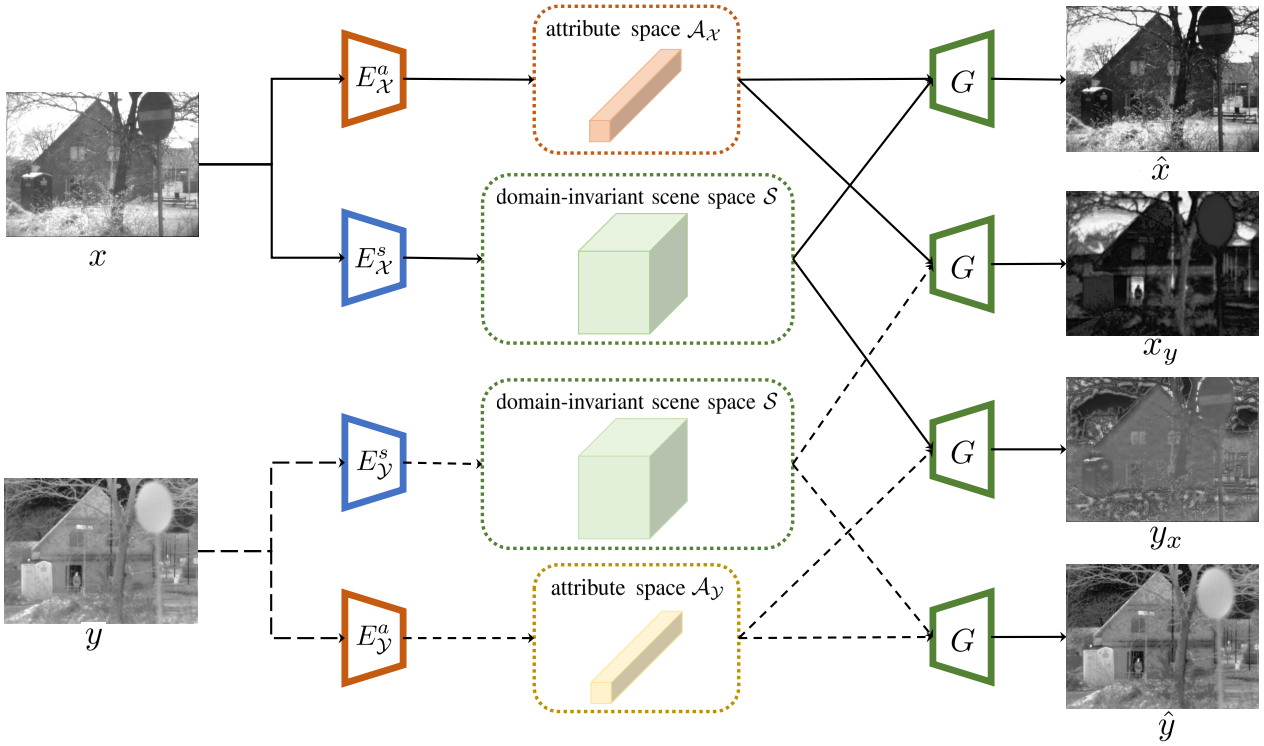

Fig. 2. Framework of disentangled representation for multimodality images. x is a source image in the visual domain $\mathcal{X}$, and y is a source image in the visual domain $\mathcal{Y}$. $\{E^s_\mathcal{X}$ and $E^a_\mathcal{X}\}$ denote the scene and attribute encoders for $\mathcal{X}$, and $\{E^s_\mathcal{Y}$ and $E^a_\mathcal{Y}\}$ are the encoders for $\mathcal{Y}$. G is the generator network. $\hat{x}$ and $\hat{y}$ are the reconstructed x and y, respectively. x_y is the fake x that is generated with the scene information of y and the attribute information of x. y_x is the fake y of which the scene information is from x and the attribute information is from y.

features from a blurred image. Wang *et al.* [30] proposed an efficient disentangled representation of disentangled features to solve the problem of cross-domain face presentation attack detection. In this article, as the VIS and IR images belong to two visual domains, we are devoted to disentangling the source images into two spaces: a domain-invariant scene space and a domain-specific attribute space. Thus, each source image can be represented with the domain-invariant scene information and the domain-specific attributes.

III. Proposed Method

In this section, we provide the problem formulation as the disentangled representation of scene and attribute space, the design of loss functions, the description of the network architectures, and the fusion block. In the end, implementations are provided in detail, including the publicly available data sets, training details, and the settings of hyperparameters.

A. Disentangle Scene and Attribute Representations

Given a VIS image x and an IR image y that belong to two visual domains $\mathcal{X} \subset \mathbb{R}^{H\times W}$ and $\mathcal{Y} \subset \mathbb{R}^{H\times W}$, respectively, our primary goal is to split the source images into a shared, domain-invariant scene space $\mathcal{S}$ and a specific attribute space $\mathcal{A}$. Given that the attribute space is unique to each domain, we denote the attribute space of $\mathcal{X}$ and $\mathcal{Y}$ as $\mathcal{A}_\mathcal{X}$ and $\mathcal{A}_\mathcal{Y}$, respectively. As the scene information is represented in different ways in the VIS and IR images, the mapping $\mathcal{X} \to \mathcal{S}$ and $\mathcal{Y} \to \mathcal{S}$ cannot be realized in the same way. In other words, the scene information cannot be extracted from x and y through the same function/parameters. Therefore, we design two scene encoders $\{E^s_\mathcal{X} : \mathcal{X} \to \mathcal{S}, E^s_\mathcal{Y} : \mathcal{Y} \to \mathcal{S}\}$, as shown in Fig. 2. These two encoders share the same network architecture but not the same weights. Besides, as the two modality attributes vary greatly, we design two attribute encoders $\{E^a_\mathcal{X}, E^a_\mathcal{Y}\}$ to learn $\mathcal{X} \to \mathcal{A}_\mathcal{X}$ and $\mathcal{Y} \to \mathcal{A}_\mathcal{Y}$, respectively.

Considering that the scene information is directly related to the space and location, the scene representation is presented in the form of feature maps, as shown in Fig. 1, whereas the attribute is related to the sensor modality and is not expected to carry the scene information. Thus, the form of vector is more suitable for the attribute information than feature maps. For the source image x, the scene features s_x and the attribute vector a_x can be encoded as

$$\{s_x, a_x\} = \left\{E^s_\mathcal{X}(x), E^a_\mathcal{X}(x)\right\}, \quad s_x \in \mathcal{S},\ a_x \in \mathcal{A}_\mathcal{X}. \tag{1}$$

Similarly, those of the source image y can be represented as

$$\{s_y, a_y\} = \left\{E^s_\mathcal{Y}(y), E^a_\mathcal{Y}(y)\right\}, \quad s_y \in \mathcal{S},\ a_y \in \mathcal{A}_\mathcal{Y}. \tag{2}$$

To achieve representation disentanglement, we perform three strategies. First, we share the weights of the last layer of $E^s_\mathcal{X}$ and $E^s_\mathcal{Y}$. In this way, the scene features of images in two domains can be embedded into a common space. However, the way of sharing the weights of the high-level layer cannot guarantee that the scene encoders encode the same information from two different domains. Thus, second, a constraint on the

scene features is performed, which makes $E^s_{\mathcal{X}}$ and $E^s_{\mathcal{Y}}$ encode the same scene features from two domains. Third, to suppress the scene information from the attribute space, we perform a constraint on the distribution of attribute vectors a_x and a_y. Thus, the attribute encoders will not encode the scene-related information.

Then, in order to make these two types of information capable of representing the source images, it should be possible to map the spaces $\mathcal{S}$ and $\mathcal{A}$ back to the original visual domains. Therefore, we adopt a generator network G to learn the inverse mapping. Considering that $\mathcal{A}_{\mathcal{X}}$ and $\mathcal{A}_{\mathcal{Y}}$ are discrepant for the generator and considering the subsequent fusion process, $\{\mathcal{S}, \mathcal{A}_{\mathcal{X}}\} \rightarrow \mathcal{X}$ and $\{\mathcal{S}, \mathcal{A}_{\mathcal{Y}}\} \rightarrow \mathcal{Y}$ share the same generator. The generator is expected to have two capacities.

On the one hand, the original source image is expected to be reconstructed conditioned on the scene and attribute representations disentangled from it. Specifically, conditioned on $\{s_x, a_x\}$ and $\{s_y, a_y\}$, the reconstructed images can be defined as

$$\hat{x} = G(s_x, a_x), \quad \hat{y} = G(s_y, a_y) \tag{3}$$

where $\hat{x}$ and $\hat{y}$ should be similar to the original x and y, respectively.

On the other hand, $\mathcal{S}$ is expected to capture the information across domains $\mathcal{X}$ and $\mathcal{Y}$, while $\mathcal{A}_{\mathcal{X}}$ and $\mathcal{A}_{\mathcal{Y}}$ should capture the domain-specific attributes without carrying the domain-invariant scene-related cues. Given that x and y are the descriptions of the same scene, s_x and s_y are supposed to be similar. Thus, given different attribute vectors, the images generated by G is supposed to be the same as those original images from which the attribute vectors are extracted from. For instance, conditioned on the s_x and a_y, G performs the translation as

$$y_x = G(s_x, a_y) \tag{4}$$

where y_x is a y-like image transformed from x with the attribute vector of y, as shown in Fig. 2. y_x and y belong to the same domain $\mathcal{Y}$. Because there are paired source images in the image fusion problem, y_x and y should maintain the pixel-level consistency. Similarly, the transformed x-like image can be defined as

$$x_y = G(s_y, a_x). \tag{5}$$

B. Loss Functions

In Section III-A, we give an intuitive description of the constraints that should be performed. In this section, we give clear definitions, i.e., the loss functions of the encoders and the generator.

1) Scene Feature Consistency Loss: Given that x and y are the descriptions of the same scene, the scene features of them are supposed to be similar. Thus, a scene feature consistency loss is defined on s_x and s_y as

$$\mathcal{L}_{\text{scene}} = \|s_x - s_y\|_1 \tag{6}$$

where $\|\cdot\|_1$ denotes the l_1-norm. Besides, the Frobenius-norm is a natural choice to constrain the consistency between feature maps. However, the l_1-norm is more suitable for this problem, and the reason is as follows. As the imaging principles of infrared and visible sensors are different, the scene information in these two types of source images cannot be exactly the same. For example, as shown in the third column in Fig. 10, when there are soldiers behind the smoke, the visible sensor cannot capture the information about them. Naturally, the information about them will not be decomposed and appear in the scene features. Conversely, the information about them is clearly presented in the infrared image. As they belong to the captured scene, the disentangled scene representation will contain the information scene. This is the significant difference between the scene representations, but it accounts for a small proportion of the scene. We expect that most proportion of the scene representation must be the same and give a certain tolerance for this special case. In other words, we expect the differences between the scene representations to be sparse. Therefore, compared with the Frobenius-norm, the l_1-norm is more suitable for this problem.

2) Attribute Distribution Loss: Based on the disentangled representation, we expect to suppress the scene information from the attribute space as much as possible. The attribute representation is expected to be as close as a prior Gaussian distribution. It has been shown in [31] that the KL term encourages disentanglement. It was suggested that the stronger pressure for the posterior match the factorized unit Gaussian prior puts extra constraints on the implicit capacity of the latent bottleneck [32]. To achieve this goal, we perform a constraint on the distribution of attribute vectors a_x and a_y by measuring the KL divergence between their distribution and a prior Gaussian distribution

$$\mathcal{L}_{\text{attr}} = \mathbb{E}[D_{\text{KL}}((a_x)\|N(0,1))] + \mathbb{E}[D_{\text{KL}}((a_y)\|N(0,1))]. \tag{7}$$

3) Self-Reconstruction Loss: The original source image is expected to be reconstructed conditioned on the scene and attribute representations disentangled from it. That is, the generator G should be able to decode the scene features and the attribute vector back to the original source image. Thus, we perform a self-reconstruction loss to make the reconstructed images achieve high fidelity with the original ones. The self-reconstruction loss is specifically defined as

$$\mathcal{L}_{\text{recon}} = \|x - \hat{x}\|_1 + \|y - \hat{y}\|_1. \tag{8}$$

4) Domain-Translation Loss: The transformed images are generated conditioned on the scene features of one source image and the attribute vectors of the other source image, which are defined as $\{x_y, y_x\} = \{G(s_y, a_x), G(s_x, a_y)\}$. Given that x and y are paired source images in the image fusion problem, y is the ideal transformed image in the domain $\mathcal{Y}$ of x. Similarly, x is the desired result of x_y. Thus, it is possible to perform a pixel-level constraint on the transformed images, which is defined as follows:

$$\mathcal{L}^{\text{domain}}_{\text{tran}} = \|x - x_y\|_1 + \|y - y_x\|_1. \tag{9}$$

Thus, the full loss function is defined as

$$\mathcal{L} = \mathcal{L}_{\text{scene}} + w_{\text{attr}}\mathcal{L}_{\text{attr}} + w_{\text{recon}}\mathcal{L}_{\text{recon}} + w_{\text{tran}}\mathcal{L}^{\text{domain}}_{\text{tran}} \tag{10}$$

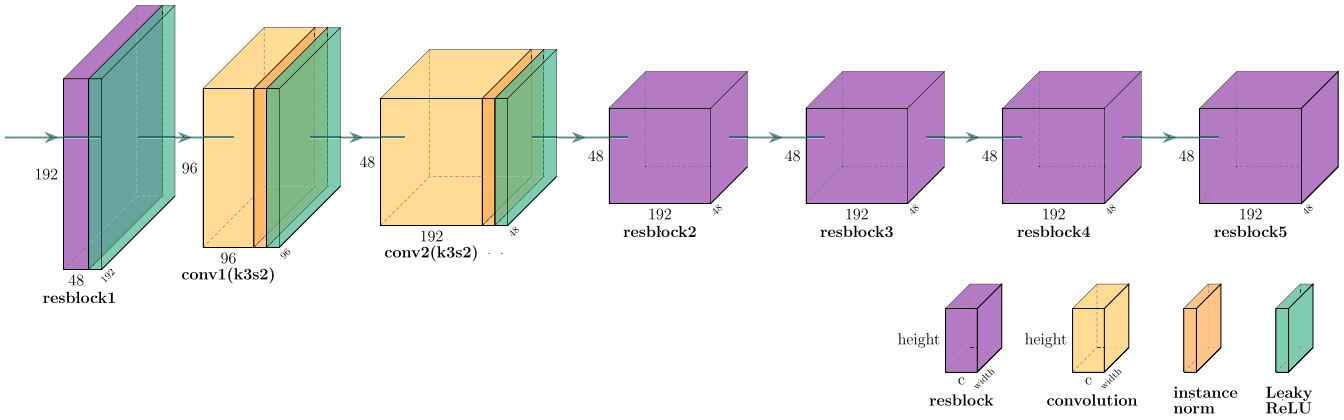

Fig. 3. Architecture of the scene encoders $\{E^s_{\mathcal{X}}, E^s_{\mathcal{Y}}\}$. c denotes the number of channels. In convolution layers, kmsn denotes that the kernel size is m and the stride is n.

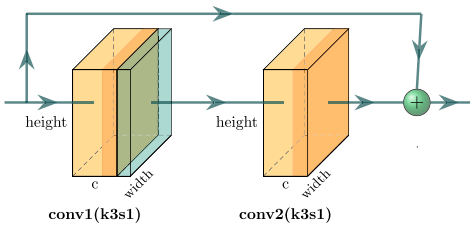

Fig. 4. Architecture of resblock.

where w_{attr}, w_{recon}, and w_{tran} are hyperparameters that control the tradeoff of each term. The parameters in the four encoders $\{E^s_{\mathcal{X}}, E^a_{\mathcal{X}}, E^s_{\mathcal{Y}}, E^a_{\mathcal{Y}}\}$ and the generator G are optimized by minimizing $\mathcal{L}$ defined in (10).

C. Network Architecture

1) Scene Encoders: The network architecture of two scene encoders $\{E^s_{\mathcal{X}}$ and $E^s_{\mathcal{Y}}\}$ is shown in Fig. 3. It consists of seven layers, including five residual blocks and two convolution layers. The residual block is applied to alleviate the vanishing gradient and degradation problem through the direct connection between the input and output [33]. The specific architecture of the residual block is shown in Fig. 4. The activation function is Leaky ReLU.

It is worth noting that, after the convolution layer, we employ the instance normalization [34] as it performs style normalization by normalizing feature statistics. These feature statistics have been found to carry the style information of an image [35], i.e., the attribute information in our method. Different from the batch normalization [36] that normalizes the mean and standard deviation (SD) based on minibatch statistics for each individual feature channel, the instance normalization computes the mean and SD across spatial dimensions independently for not only each feature channel but also each sample. Mathematically, given an input batch $u \in \mathbb{R}^{N\times H\times W\times C}$, the normalized u is computed as

$$\text{IN}(u) = \gamma\left(\frac{u-\bar{u}}{\sigma}\right) + \beta \tag{11}$$

where N, H, W, and C denote the batch size, height, width, and the number of channels, respectively. γ and β are the affine parameters. $\bar{u}$ and $\sigma \in \mathbb{R}^{N\times C}$ are the mean and SD computed across spatial dimensions that are mathematically defined as

$$\bar{u}_{nc} = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} u_{nhwc} \tag{12}$$

$$\sigma_{nc} = \sqrt{\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}(u_{nhwc}-\bar{u}_{nc})^2 + \epsilon} \tag{13}$$

where n, h, w, and c denote the data in the nth batch, the hth row, the wth column, and the cth channel. ϵ is a small value used to maintain stability.

In addition, based on the assumption that the scene features extracted from domains $\mathcal{X}$ and $\mathcal{Y}$ share the same scene space $\mathcal{S}$, we share the weight of the last residual block in the scene encoders. More concretely, $E^s_{\mathcal{X}}$ and $E^s_{\mathcal{Y}}$ share the weight of resblock5 in Fig. 3. In this way, the scene representation is forced to be mapped into a common scene space.

2) Attribute Encoders: As shown in Fig. 5, the first five layers of the attribute encoders are traditional convolution layers with the kernel size set as 5×5 and the stride set as 2. Then, through the global average pooling layer across spatial dimensions, the attribute information is mapped into a vector. Through the sixth convolution layer, the final z-dimensional attribute vector is obtained. To make $\mathcal{A}_x$ and $\mathcal{A}_y$ two distinctive attribute spaces for the generator, we give a bias to the attribute vectors in $\mathcal{A}_y$ make them distinguished from those in $\mathcal{A}_x$.

3) Generator: The network architecture of the generator G is shown in Fig. 6. For the scene features, they are first passed through a residual block. For the attribute vector, it is tiled into the same width and height as the scene features. The output of the first residual block and that of the tile layer are concatenated and fed into the subsequent residual blocks. Then, two deconvolution layers are used for upsampling the feature maps. It is worth noting that the spatial resolution of the scene features is reduced to a quarter of the original image, so many high-quality texture details are lost. Inspired by U-net [37], to preserve the lost information, the output of the first residual block in the scene encoder, i.e., low-level features, is also used as a part of the scene information. It is

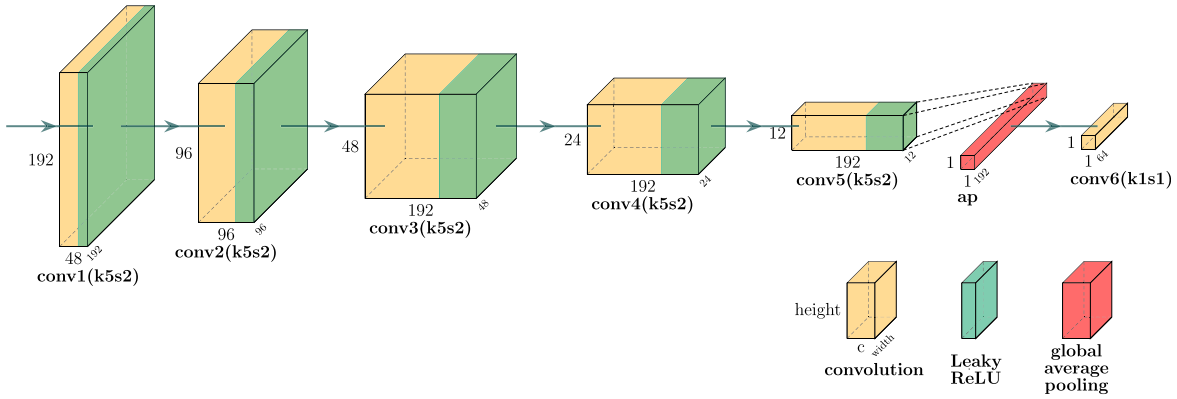

Fig. 5. Architecture of the attribute encoders $\{E^a_{\mathcal{X}}$ and $E^a_{\mathcal{Y}}\}$.

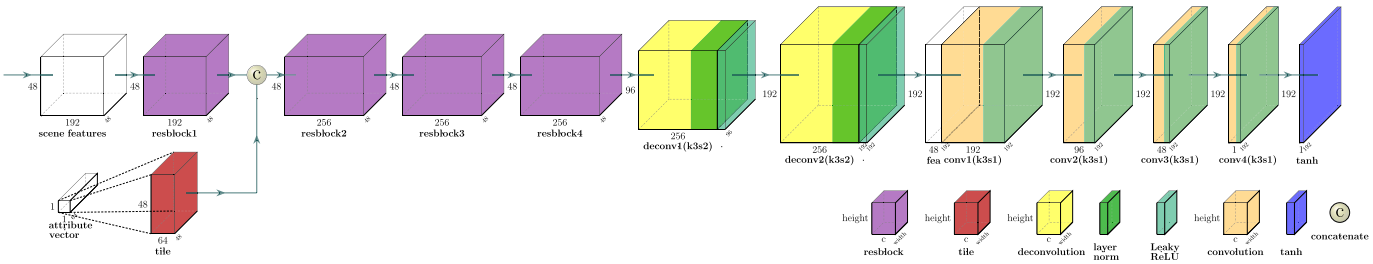

Fig. 6. Architecture of the generator G ("fea" in the conv1 is the features extracted by the scene encoder through the resblock1 in Fig. 3 with high spatial resolution).

concatenated with the output of the second deconvolution layer and fed into the first convolution layer in the generator. After being passed through the following four convolution layers, the channels of the feature maps are reduced gradually into the channel of the original image. Through the tanh activation function, the reconstructed images are generated.

It is worth noting that since the instance normalization unifies the style (attribute) of images, the introduction of instance normalization is not conducive to the generation of the image with various styles. Thus, the instance normalization is not applied after the convolution layers in the generator.

D. Fusion Block

With the pretrained encoders and generator, the fusion process is performed on the scene space $\mathcal{S}$ and the attribute space $\mathcal{A}$ individually according to the disentangled representations.

1) Scene Representation Fusion: The scene features s_x and s_y are assumed to share the same scene. Besides, based on the strategy of weight sharing between the last residual blocks of two scene encoders and the scene feature consistency loss defined in (6), s_x and s_y are mapped into a common scene space. Thus, we perform the average strategy to obtain the fused scene features as

$$s_f = \frac{s_x + s_y}{2}. \tag{14}$$

2) Attribute Representation Fusion: For attribute vectors, we directly apply the addition fusion strategy. The fused attribute vector is defined as

$$a_f = \lambda a_x + (1 - \lambda) a_y \tag{15}$$

where λ is a hyperparameter between 0 and 1, which is used to modulate the presented attribute of the fusion result. Specifically, when $\lambda = 0$, the fusion result looks like an image belonging to the visual domain $\mathcal{Y}$; while $\lambda = 1$, the result seems to be similar to the images in the domain $\mathcal{X}$. For the subsequent various application targets, the fusion result can be modulated to present different attributes by setting different λ's. Because the scene features have been disentangled, fused, and fixed as s_f for the fusion result, the different setting of λ has little effect on the scene information distortion. The qualitative results about λ will be given later in Section IV-A.

Finally, the fused scene features and fused attribute vectors are fed into the pretrained generator to produce the final fused image f, which can be represented as

$$f = G(s_f, a_f). \tag{16}$$

E. Implementations

1) Data: Both the training and test data are from publicly available data sets. Considering that the networks need a large quantity of data to train the parameters, we use the image pairs in the *RoadScene* data set[1] provided by Xu *et al.* [38], [39] to establish the training data. The reason is that *RoadScene* contains 221 aligned VIS and IR image pairs that contain rich scenes, such as pedestrians, roads, and vehicles. These pairs are highly representative scenes from the FLIR video.[2] The background thermal noise is preprocessed, and image pairs are aligned in this data set. It solves the problems in existing

[1] Available at https://github.com/hanna-xu/RoadScene.

[2] Available at https://www.flir.com/oem/adas/adas-data set-form/.

data sets, such as few image pairs, low spatial resolution, and extreme lack of detailed information in the infrared image.

We select 150 image pairs from the *RoadScene* data set to establish the training data set. As the VIS images in *RoadScene* are RGB images, we first transfer them into a single channel (Y channel in the YCbCr color space). Then, we use a window of 192×192 to slide from the upper left corner to the lower right corner of these 150 single-channel image pairs with an overlap of 32. Then, these image pairs are cropped into 2160 patch pairs with the size of 192×192. These patches are used as the training data to train our networks. The way and results of fusing RGB VIS images and single-channel IR images will be discussed and shown later in Section IV-B.

In the testing phase, to validate the generalization ability of the algorithm, we test our method on two VIS and IR data sets, including the *RoadScene* and *TNO Human Factors*[3] data sets. *TNO* is a standard VIS-IR image pair data set. It contains about 50 multispectral nighttime imagery of many military relevant scenarios registered with different multiband camera systems. It contains scenarios, such as tank, sand path, bunker, lake, bench, and helicopter. Both the VIS and IR images in *TNO* are single-channel images.

2) Training Details: The dimension of the attribute vectors, i.e., z, is fixed as 96. We set $w_{\text{attr}} = 0.001$, $w_{\text{recon}} = 10$, and $w_{\text{tran}} = 40$. λ is empirically set as 0.3. The epoch is 5, and the batch size is 4. The parameters are updated by the Adam optimizer with the learning rate set as 0.0001 with exponential decay. The proposed algorithm is implemented in TensorFlow. Experiments are performed on NVIDIA Geforce GTX Titan X GPU and 3.4-GHz Intel Core i5-7500 CPU. It takes about 70 min to train the proposed network. As there is no fully connected layer in our network, the proposed method is generalizable to images of any dimension. However, as there are two deconvolution layers, the width and height of source images are preferably multiples of 4. Otherwise, the size of the generated image will deviate that of source images by one to three points.

IV. Experimental Results

A. Qualitative Evaluation

To validate the effectiveness of our proposed method, we compare our DRF with five state-of-the-art methods, including three traditional methods, i.e, GTF [16], NSCT [40], FPDE [41], and two deep learning-based methods, i.e., DenseFuse [42] and FusionGAN [18]. Both the qualitative and quantitative comparisons are preformed on the publicly available *TNO* and *RoadScene* data sets.

The qualitative comparison results on the *TNO* data set are shown in Fig. 7, and those on the *RoadScene* data set are shown in Fig. 8. As the target of image fusion is to extract the vital and complementary information from source images and use it to generate a single fused image, the quality evaluation mainly focuses on whether the vital/complementary information is preserved or degraded. From the perspective of subsequent applications (e.g., surveillance and vehicle navigation and monitoring), we also qualitatively evaluate whether the thermal targets are prominent or not. Compared with the five competitors, the results of our DRF have three distinctive advantages: 1) our results can retain the information from both of the source images at the same time; 2) when there is little information in one of the source images, the fusion result is not degraded in our method; and 3) our method can maintain the thermal radiation property of the scene.

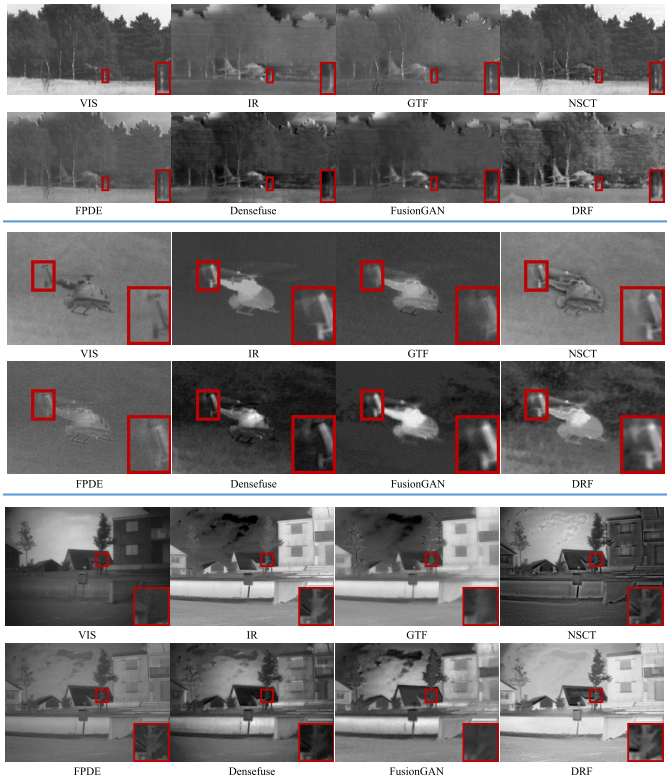

Fig. 7. Qualitative comparison of our DRF with five state-of-the-art methods on three typical VIS-IR image pairs in the *TNO* data set. In each group of results, from top to bottom (from left to right): visible and infrared image pair, fusion results of GTF [16], NSCT [40], FPDE [41], Densefuse [42], FusionGAN [18], and our DRF.

The first advantage can be seen from both the first group of results in Figs. 7 and 8. As shown in the results on these image pairs, because the type of information to be retained in each source image is specified, e.g., the pixel intensity distribution of the infrared images and the gradients of the visible images, the results of the competitors lose some information from the visible images or the infrared images. In our method, the information is fused through the scene features that are the intrinsic properties of the scene and have nothing to do with the way of presentation. Thus, the information from both of the source images can be retained at the same time in our method.

The second advantage can be seen from the last two groups of results in Fig. 7 and the second group of results in Fig. 8. In these image pairs, there is less information in the visible images. In the results of the competitors, the introduction of information from the visible images degrades the quality of thermal information in the infrared images. In other words,

[3]Available at https://figshare.com/articles/TNO_Image_Fusion_Dataset/1008029.

Fig. 8. Qualitative comparison of our DRF with five state-of-the-art methods on three typical VIS-IR image pairs in the *RoadScene* data set. In each group of results, from top to bottom (from left to right): visible and infrared image pair, fusion results of GTF [16], NSCT [40], FPDE [41], Densefuse [42], FusionGAN [18], and our DRF.

the introduction of low-quality information in one source image leads to the loss of distinctive information in the other source image. In our method, this problem is alleviated because the information is not fused through the surface properties.

The third advantage is that our method can maintain the thermal radiation property of the scene, which can be seen from the last group of results in Fig. 8. On this image pair, all the competitors except FPDE reverse the thermal radiation property, as shown in the red boxes. As the sign is of the high pixel value in the visible image, the competitors mistakenly highlight it as a thermal target. Besides, the results of FPDE fail to highlight the thermal targets, while our method can maintain the thermal radiation property and keep the prominence of thermal targets.

B. Qualitative Results in RGB Version

In the *RoadScene* data set, the VIS images are in the RGB version. Thus, in this section, we perform the experiment of fusing the RGB VIS image and gray IR image. To this end, the RGB VIS image is first converted from the RGB color space into the YCbCr color space. The Y channel is the luminance channel, and the Cb and Cr channels are the chrominance channels. As the structures are usually in the luminance channel, we fuse the Y channel of VIS image and an IR image. Then, the fused Y channel is concatenated with two chrominance channels and then converted into the RGB color space to obtain the final fusion result. The qualitative results are shown in Fig. 9. As shown in this figure, the fused images look like the IR images enhanced with the chrominance and texture information of the VIS images.

Fig. 9. Qualitative fusion results on the four image pairs in the *RoadScene* data set. From top to bottom: the RGB VIS images, the gray IR images, and the RGB fusion results.

C. Experiment on λ

A characteristic of our proposed method is that the scene and attribute information of the source images are fused separately. The fusion of the scene features aims to include all the objects in the scene from multiple source images, while the fusion of attribute information determines the visual presentation of the fusion result. Because the fused attribute vector depends on the value of λ, λ plays the role of modulating the visual presentation of a fused image.

To demonstrate this property of λ, the qualitative results of different λ on the five image pairs are provided in Fig. 10. With the increase in λ, the attribute of fusion results approaches from the infrared attribute to the visible attribute gradually. The prominence of thermal targets gradually decreases, but the texture details gradually become abundant. We can see that, when $\lambda = 0$, the fusion results look like infrared images because the fused attributes are all come from those of the infrared images. Thus, the fusion results belong to the infrared domain. Similarly, when $\lambda = 1$, the results belong to the visible domain and look like the visible images. As λ increases, the fusion results transform from the infrared and visible domains. The prominence of the thermal targets, which is the characteristic of the infrared domain, decreases gradually, while the texture details, which are the characteristic of the visible domain, are gradually enhanced. Through a comprehensive consideration of the prominence of the thermal targets and the richness of the details, λ is set as 0.3 for satisfactory results.

It is worth noting that even though the results are similar to the source images when $\lambda = 0$ or 1, there are still obvious differences between the results and source images, which can be obviously seen in the third column. When $\lambda = 1$, the soldiers behind the smoke are clear to be seen, while the corresponding visible image does not contain the information about soldiers due to the occlusion of the smoke. This phenomenon is caused by the fusion of the scene features, which aims to fuse the object information in the source images into the fusion result. Even though the results of $\lambda = 1$

Fig. 10. Qualitative fusion results on the five image pairs (the first three image pairs are from the *TNO* data set, and the last two image pairs are from the *RoadScene* data set). From top to bottom: infrared image, fusion results when $\lambda = 0, 0.2, 0.4, 0.6, 0.8, 1$, and visible image.

seem like the visible images, the scene information of them have included the scene information of the infrared images and, thus, show a little different compared with the visible images.

D. Quantitative Evaluation

In this section, we further perform the quantitative comparisons of our DRF with other competitors on both the *TNO* and *RoadScene* data sets. We use six metrics for evaluation. More concretely, we use the entropy (EN) and SD to evaluate the nature of the fusion results themselves. EN measures the amount of information contained in the fusion result from the perspective of information theory. A large EN means that the fused image contains much information and, thus, exhibits good performance. SD reflects the contrast and distribution with the SD of the fused image. The larger SD, the higher contrast the fused image achieves. As the attention of humans is more likely to be attracted by the area with high contrast,

TABLE I

QUANTITATIVE EVALUATION OF DIFFERENT METHODS ON 15 IMAGE PAIRS IN THE *TNO* DATA SET. MEAN AND SD OF DIFFERENT METRICS ARE SHOWN IN THIS TABLE. DRF W/O DR INDICATES THE PROPOSED METHOD WITHOUT DISENTANGLED REPRESENTATION, I.E., $w_{\text{TRAN}} = 0$ (RED: OPTIMAL, BLUE: SUBOPTIMAL, AND PINK: THIRD OPTIMAL)

	GTF[16]	NSCT [40]	FPDE [41]	Densefuse [42]	FusionGAN [18]	DRF w/o DR	DRF
EN	6.659 ± 0.662	6.915 ± 0.466	6.357 ± 0.432	6.701 ± 0.389	6.039 ± 0.426	6.811 ± 0.374	6.864 ± 0.371
SD	30.163 ± 9.829	45.857 ± 14.972	25.051 ± 6.281	27.483 ± 4.533	23.388 ± 5.498	30.634 ± 3.881	30.853 ± 6.597
PSNR	13.808 ± 1.856	13.180 ± 1.865	15.975 ± 2.388	12.105 ± 1.952	12.846 ± 2.051	14.150 ± 1.659	14.313 ± 1.600
FMI	0.876 ± 0.036	0.886 ± 0.037	0.871 ± 0.032	0.873 ± 0.039	0.866 ± 0.029	0.877 ± 0.035	0.879 ± 0.030
SSIM	0.624 ± 0.153	0.603 ± 0.109	0.647 ± 0.165	0.613 ± 0.110	0.619 ± 0.149	0.643 ± 0.116	0.641 ± 0.128
CC	0.555 ± 0.067	0.118 ± 0.166	0.443 ± 0.154	0.513 ± 0.099	0.591 ± 0.068	0.512 ± 0.156	0.519 ± 0.079

TABLE II

QUANTITATIVE EVALUATION OF DIFFERENT METHODS ON 30 IMAGE PAIRS IN THE *RoadScene* DATA SET. MEAN AND SD OF DIFFERENT METRICS ARE SHOWN IN THIS TABLE (RED: OPTIMAL, BLUE: SUBOPTIMAL, AND PINK: THIRD OPTIMAL)

	GTF[16]	NSCT [40]	FPDE [41]	Densefuse [42]	FusionGAN [18]	DRF w/o DR	DRF
EN	7.361 ± 0.287	7.074 ± 0.197	6.992 ± 0.262	7.314 ± 0.205	6.840 ± 0.303	7.244 ± 0.266	7.361 ± 0.243
SD	49.540 ± 10.233	39.120 ± 4.783	34.905 ± 6.054	45.890 ± 6.189	36.567 ± 7.440	45.365 ± 5.783	49.308 ± 7.599
PSNR	13.834 ± 2.811	12.396 ± 2.880	15.688 ± 2.744	12.753 ± 1.593	12.939 ± 2.244	12.994 ± 1.972	13.144 ± 2.937
FMI	0.865 ± 0.023	0.870 ± 0.024	0.837 ± 0.033	0.823 ± 0.028	0.841 ± 0.031	0.847 ± 0.030	0.848 ± 0.023
SSIM	0.643 ± 0.079	0.690 ± 0.075	0.716 ± 0.151	0.476 ± 0.075	0.636 ± 0.093	0.677 ± 0.074	0.651 ± 0.066
CC	0.622 ± 0.148	0.265 ± 0.318	0.640 ± 0.184	0.552 ± 0.184	0.606 ± 0.154	0.596 ± 0.154	0.614 ± 0.148

the larger SD shows that the fused image shows a better visual effect.

In addition, we use the peak signal-to-ratio (PSNR), feature mutual information (FMI) [43], structural similarity index measure (SSIM) [44], and correlation coefficient (CC) to measure the relationship between the fused and source images. PSNR reflects the similarity between the fusion result and source images and the distortion caused by the fusion process. A large PSNR indicates that the fused image is similar to the source images, and the fusion process produces little distortion. FMI is based on MI and feature information. It measures the feature information transferred from source images to the fusion result. A large FMI indicates that considerable feature information is transferred from source images into the fusion result. SSIM measures the image loss and distortion from three aspects: correlation, luminance, and contrast. A large SSIM indicates that the fusion result can achieve high structural similarities with the source images. CC measures the degree of linear correlation of the fused image and source images. The larger the CC, the more similar the fused image is to the source images and the better the fusion performance.

The quantitative results on the *TNO* data set are shown in Table I, and those on the *RoadScene* data set are shown in Table II. The mean and SD in these tables show that our DRF can achieve comparable results on the three metrics. It achieves the suboptimal results on the three metrics on the *TNO* data set, while the algorithms that exhibit the optimal results are not the same. For the *RoadScene* data set, our method achieves the best result on EN and the suboptimal result on SD. The results on these two data sets show that our results have comparable information, higher contrast, and less distortion. The reason for our suboptimal results on PSNR is that our method tries to keep more attributes of the infrared images as λ is set as 0.3 rather than 0.5. The highest result of FPDE on PSNR shows that it achieves the highest similarity between the fusion result and source images. However, PSNR is measured mainly based on the pixel intensity and without considering other factors, such as structure and contrast. Thus, when the pixel intensity of the fusion result is in the middle of those of source images, the fusion result can achieve a high value on PSNR, whereas, in this condition, the prominence of thermal targets is weakened. This phenomenon is obviously reflected in the results of FPDE. In other methods that expect to retain the prominence, including our method, their results on PSNR are not high.

On the other hand, it is worth noting that the IR images in the *RoadScene* data set contain more information and perform high quality than those in the *TNO* data set, which can be seen from the IR images in Figs. 7 and 8. Thus, the competitors show different performances on these two data sets. For instance, NSCT achieves the best results on EN and SD on the *TNO* data set, while, on the *RoadScene* data set, it shows slightly weaker results. However, the results of GTF on these metrics are contrary to those of NSCT. GTF achieves comparable results on the *RoadScene* data set but mediocre performance on the *TNO* data set. Thus, the performance of these competitors is affected by the source image quality, at least the quality of the IR images. By comparison, our method shows comparable results on both of these data sets. It shows that our method has a good generalization for images with different qualities.

E. Ablation Study

In our method, to realize the disentangled representation, we apply the domain-translation loss to transfer the images from one domain to the other domain. To validate its effectiveness, we perform the ablation study in this section where the domain-translation loss is not applied, i.e., DRF w/o disentanglement. In this condition, the whole network is no

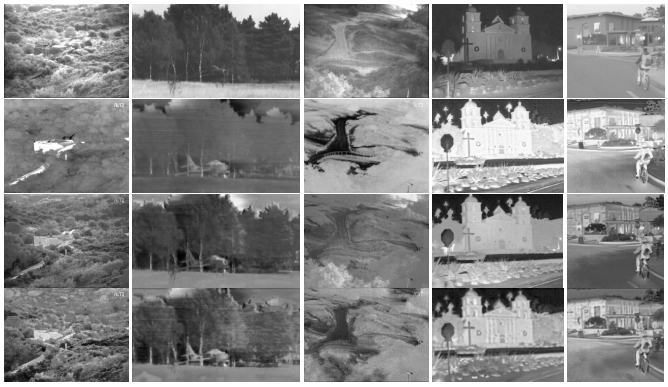

Fig. 11. Qualitative ablation study of the domain-translation loss (the first three image pairs are from the *TNO* data set, and the last two image pairs are from the *RoadScene* data set). From top to bottom: visible and infrared images, fusion results without the domain-translation loss ($w_{\text{tran}} = 0$), and results by applying the domain-translation loss (DRF).

longer formulated for disentangled representation. Instead, the four encoders are trained merely as the feature extraction, and the generator is used for reconstruction. The outputs of the encoders do not have the physical meaning of scene- or attribute-related information.

The qualitative results on five image pairs are shown in Fig. 11. As shown in this figure, with the same fusion strategy, by removing the domain-translation loss ($w_{\text{tran}} = 0$), the prominence of the thermal targets is weakened. In other words, more information of the visible image/less information of the infrared image is retained in the fused image. The quantitative results are shown in Tables I and II. In almost all metrics on the two data sets, DRF shows superior performance than DRF w/o disentanglement. The reason is that, by applying this loss, we rely on λ (the weight of different attribute representations) to control the representation form of the fusion result, whereas, when this loss function is removed, λ and the outputs of the attribute encoders lose their functions. The fusion process is mainly realized by averaging the extracted feature maps. In this way, the fusion results are similar to those of Densefuse and fail to highlight the thermal targets.

A limitation of the proposed method is that it is designed to fuse aligned visible and infrared images. When the source images suffer from some alignment errors, our method will fail to handle this situation and align them. Thus, the fusion result may show obvious scene deviation.

V. Conclusion

In this article, we propose a novel decomposition method for infrared and visible image fusion by applying disentangled representation, named DRF. According to the imaging principle, we perform the decomposition depending on the source of information in the visible and infrared images. More concretely, we disentangle the images into the scene- and sensor modality (attribute)-related representations through the corresponding encoders, respectively. Then, different strategies are applied for the fusion of these different types of representations. Finally, the fused representations are fed into a pretrained generator to generate the fusion result. The qualitative and quantitative experiments on the publicly available *TNO* and *RoadScene* data sets demonstrate the comparable performance of our DRF over the state of the art, in terms of both visual effect and quantitative metrics. The qualitative results by applying different hyperparameters also show that our method can adjust the characteristics of the fusion results according to different application targets.

Furthermore, as in medical image fusion, the images are captured from the same part of the human body but with different imaging systems (usually structural and functional systems). Thus, the medical image fusion is similar to VIF in essence. These medical images can also be disentangled into the body-related information and imaging system-related representation. Also, in multiexposure image fusion, the multiexposure images are captured with the same scene but multiple exposure settings. The scene-related representation and exposure setting-related representation can also be disentangled from the source images. Thus, in our future work, we will apply the disentangled representation to multimodality medical image fusion and multiexposure image fusion.

References

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

[2] H. Li, X.-J. Wu, and T. Durrani, "NestFuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 12, pp. 9645–9656, Dec. 2020.

[3] J. Ma and Y. Zhou, "Infrared and visible image fusion via gradientlet filter," *Comput. Vis. Image Understand.*, vols. 197–198, Aug. 2020, Art. no. 103016.

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

[5] L. Jian, X. Yang, Z. Liu, G. Jeon, M. Gao, and D. Chisholm, "Sedrfuse: A symmetric encoder–decoder with residual block network for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–15, 2020.

[6] B. Yang, S. Li, and F. Sun, "Image fusion using nonsubsampled contourlet transform," in *Proc. Int. Conf. Image Graph.*, 2007, pp. 719–724.

[7] W. Gan *et al.*, "Infrared and visible image fusion with the use of multi-scale edge-preserving decomposition and guided image filter," *Infr. Phys. Technol.*, vol. 72, pp. 37–51, Sep. 2015.

[8] J. Chen, X. Li, L. Luo, X. Mei, and J. Ma, "Infrared and visible image fusion based on target-enhanced multiscale transform decomposition," *Inf. Sci.*, vol. 508, pp. 64–78, Jan. 2020.

[9] Y. Yang, Y. Zhang, S. Huang, Y. Zuo, and J. Sun, "Infrared and visible image fusion using visual saliency sparse representation and detail injection model," *IEEE Trans. Instrum. Meas.*, vol. 70, 2021, Art. no. 5001715.

[10] J. Wang, J. Peng, X. Feng, G. He, and J. Fan, "Fusion method for infrared and visible images by using non-negative sparse representation," *Infr. Phys. Technol.*, vol. 67, pp. 477–489, Nov. 2014.

[11] M. Yin, P. Duan, W. Liu, and X. Liang, "A novel infrared and visible image fusion algorithm based on shift-invariant dual-tree complex shearlet transform and sparse representation," *Neurocomputing*, vol. 226, pp. 182–191, Feb. 2017.

[12] Z. Zhu, H. Yin, Y. Chai, Y. Li, and G. Qi, "A novel multi-modality image fusion method based on image decomposition and sparse representation," *Inf. Sci.*, vol. 432, pp. 516–529, Mar. 2018.

[13] H. Li and X.-J. Wu, "Infrared and visible image fusion using latent low-rank representation," 2018, *arXiv:1804.08992*. [Online]. Available: http://arxiv.org/abs/1804.08992

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

[15] B. Cheng, L. Jin, and G. Li, "General fusion method for infrared and visual images via latent low-rank representation and local non-subsampled shearlet transform," *Infr. Phys. Technol.*, vol. 92, pp. 68–77, Aug. 2018.

[16] J. Ma, C. Chen, C. Li, and J. Huang, "Infrared and visible image fusion via gradient transfer and total variation minimization," *Inf. Fusion*, vol. 31, pp. 100–109, Sep. 2016.
[17] Q. Du, H. Xu, Y. Ma, J. Huang, and F. Fan, "Fusing infrared and visible images of different resolutions via total variation model," *Sensors*, vol. 18, no. 11, p. 3827, Nov. 2018.
[18] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Inf. Fusion*, vol. 48, pp. 11–26, Aug. 2019.
[19] J. Ma *et al.*, "Infrared and visible image fusion via detail preserving adversarial learning," *Inf. Fusion*, vol. 54, pp. 85–98, Feb. 2020.
[20] J. Ma, H. Xu, J. Jiang, X. Mei, and X.-P. Zhang, "DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4980–4995, 2020.
[21] J. Han, E. J. Pauwels, and P. de Zeeuw, "Fast saliency-aware multi-modality image fusion," *Neurocomputing*, vol. 111, pp. 70–80, Jul. 2013.
[22] H. Li, X.-J. Wu, and J. Kittler, "Infrared and visible image fusion using a deep learning framework," in *Proc. Int. Conf. Pattern Recognit.*, 2018, pp. 2705–2710.
[23] I. Shopovska, L. Jovanov, and W. Philips, "Deep visible and thermal image fusion for enhanced pedestrian visibility," *Sensors*, vol. 19, no. 17, p. 3727, Aug. 2019.
[24] Q. Zhang, N. Huang, L. Yao, D. Zhang, C. Shan, and J. Han, "RGB-T salient object detection via fusing multi-level CNN features," *IEEE Trans. Image Process.*, vol. 29, pp. 3321–3335, 2020.
[25] J. Li, H. Huo, C. Li, R. Wang, and Q. Feng, "AttentionFGAN: Infrared and visible image fusion using attention-based generative adversarial networks," *IEEE Trans. Multimedia*, early access, May 28, 2020, doi: 10.1109/TMM.2020.2997127.
[26] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1415–1424.
[27] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 35–51.
[28] H.-Y. Lee *et al.*, "DRIT++: Diverse image-to-image translation via disentangled representations," *Int. J. Comput. Vis.*, vol. 128, nos. 10–11, pp. 2402–2417, Nov. 2020.
[29] B. Lu, J.-C. Chen, and R. Chellappa, "Unsupervised domain-specific deblurring via disentangled representations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 10225–10234.
[30] G. Wang, H. Han, S. Shan, and X. Chen, "Cross-domain face presentation attack detection via multi-domain disentangled representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 6678–6687.
[31] R. T. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud, "Isolating sources of disentanglement in variational autoencoders," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 2610–2620.
[32] I. Higgins *et al.*, "beta-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. Int. Conf. Learn. Represent.*, 2017.
[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[34] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016, *arXiv:1607.08022*. [Online]. Available: http://arxiv.org/abs/1607.08022
[35] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1501–1510.
[36] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167
[37] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent*, 2015, pp. 234–241.
[38] H. Xu, J. Ma, Z. Le, J. Jiang, and X. Guo, "Fusiondn: A unified densely connected network for image fusion," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12484–12491.
[39] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jul. 28, 2020, doi: 10.1109/TPAMI.2020.3012548.
[40] H. Li, H. Qiu, Z. Yu, and Y. Zhang, "Infrared and visible image fusion scheme based on NSCT and low-level visual features," *Infr. Phys. Technol.*, vol. 76, pp. 174–184, May 2016.
[41] D. P. Bavirisetti, G. Xiao, and G. Liu, "Multi-sensor image fusion based on fourth order partial differential equations," in *Proc. Int. Conf. Inf. Fusion*, 2017, pp. 1–9.
[42] H. Li and X.-J. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2614–2623, May 2019.
[43] M. B. A. Haghighat, A. Aghagolzadeh, and H. Seyedarabi, "A non-reference image fusion metric based on mutual information of image features," *Comput. Electr. Eng.*, vol. 37, no. 5, pp. 744–756, Sep. 2011.
[44] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

Han Xu received the B.S. degree from the Electronic Information School, Wuhan University, Wuhan, China, in 2018, where she is currently pursuing the Ph.D. degree with the Multi-Spectral Vision Processing Lab, Electronic Information School.

She has first-authored several refereed journal articles and conference papers, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, AAAI, and IJCAI. Her current research interests include computer vision and image processing.

Xinya Wang received the B.S. degree from the Electronic Information School, Wuhan University, Wuhan, China, in 2018, where she is currently pursuing the Ph.D. degree with the Multi-Spectral Vision Processing Lab.

Her current research interests include neural networks, machine learning, and image processing.

Jiayi Ma (Member, IEEE) received the B.S. degree in information and computing science and the Ph.D. degree in control science and engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2008 and 2014, respectively.

He is currently a Professor with the Electronic Information School, Wuhan University, Wuhan. He has authored or coauthored more than 140 refereed journal articles and conference papers, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, *International Journal of Computer Vision* (IJCV), CVPR, ICCV, and ECCV. His research interests include computer vision, machine learning, and pattern recognition.

Dr. Ma has been identified in the 2020 and 2019 Highly Cited Researcher lists from the Web of Science Group. He is also an Area Editor of *Information Fusion*, an Editorial Board Member of *Neurocomputing* and *Entropy*, and a Guest Editor of *Remote Sensing*.