Deep Retinex Fusion for Adaptive Infrared and Visible Image Super-resolution Fusion

Yuanjie Gu, Zhibo Xiao, Hailun Wang, Cheng Liu, and Shouyu Wang

Abstract—Convolutional neural networks have turned into an illustrious tool for image fusion and super-resolution. However, their excellent performance cannot work without large fixed-paired datasets; and additionally, these high-demanded ground truth data always cannot be obtained easily in fusion tasks. In this study, we show that, the structures of generative networks capture a great deal of image feature priors, and then these priors are sufficient to reconstruct high-quality fused super-resolution result using only low-resolution inputs. By this way, we propose a novel self-supervised dataset-free method for adaptive infrared (IR) and visible (VIS) image super-resolution fusion named Deep Retinex Fusion (DRF). The key idea of DRF is first generating component priors which are disentangled from physical model using our designed generative networks ZipperNet, LightingNet and AdjustingNet, then combining these priors which captured by networks via adaptive fusion loss functions based on Retinex theory, and finally reconstructing the super-resolution fusion results. Furthermore, in order to verify the effectiveness of our reported DRF, both qualitative and quantitative experiments via comparing with other state-of-the-art methods are performed using different test sets. These results prove that, comparing with large datasets trained methods, DRF which works without any dataset achieves the best super-resolution fusion performance; and more importantly, DRF can adaptively balance IR and VIS information and has good noise immunity. DRF codes are open source available at https://github.com/GuYuanjie/Deep-Retinex-fusion.

Index Terms—Infrared and visible image fusion, super-resolution, dataset-free self-supervised disentangled learning, Retinex theory.

I. INTRODUCTION

INFRAEED imaging focuses on wavelengths between 8-14 µm [1] which indicating heat radiated from objects, therefore it can highlight thermal radiation objects even under poor lighting conditions or in the case of severe sheltering. Therefore, infrared imaging has many applications, such as in military reconnaissance [2], safety inspection [3], non-destructive testing [4], unmanned driving [5] and so on [6], [7], [8]. However, infrared imaging still suffers from low resolution, contrast, and signal-to-noise ratio. Compared to infrared imaging, visible light imaging focusing on wavelengths between 380-760 nm can capture reflected light often guaranteeing abundant texture details of targets. However, visible light imaging is invalid in low lighting or sheltering conditions. Therefore, IR and VIS image fusion has more extensive applications, but it is still a challenging work because conventional fusion methods cannot balance the relation of dynamic range, edge enhancement and lightness constancy.

However, deep learning can partially solve above problems to some extent, thus an increasing number of studies on deep learning based IR and VIS image fusion have been reported. In fact, it is impossible to obtain real fused image directly. In other words, the ground truth (label) as the essential element in supervised deep learning cannot be obtained. Therefore, most of these works are self-supervised (unsupervised), and can be roughly divided into two categories with [9], [10], [11] and without [12], [13] training phase, respectively. For methods with training phase, most of them are CNN based [9], [11], [14] and GAN based [10], [15], [16]. Some CNN based methods are not end-to-end framework: they use CNN to extract deep features, while adopt conventional fusion methods often relying on complicated rules for fusion. Other CNN based methods are end-to-end framework, but due to the lack of ground truth, these methods solve it through separate design on loss function or network, inevitably leading to incomplete comprehensive performance. Additionally, most of GAN methods is to force the network generate the fused image similar to both IR and VIS source images, while fusion ratio between IR and VIS images is not equal or even not linear. Besides, for methods without training phase, they first use pre-trained models to extract multi-level deep features, and then reconstruct fused image according to these extracted features. However, as most of pre-trained models are trained for various tasks, they are not fit for the specific IR and VIS image fusion.

In order to overcome these drawbacks in IR and VIS image fusion, we propose a self-supervised disentangled learning method based on Retinex theory [17] named Deep Retinex Fusion (DRF). According to Retinex theory, the captured image comes from the interaction between the light (radiation) source L and the object R formulized as (1).

\[
I(x, y) = R(x, y) \cdot L(x, y)
\]

(1)

Retinex based on lightness consistency is different from conventional linear and nonlinear methods (like power-law function, gamma function and histogram equalization), which can only enhance one type of image features. It can balance dynamic range compression, edge enhancement and lightness constancy, thus enabling adaptive en-hancement of various types of images. Inspired by the above theory, we implement a novel deep learning version of Retinex named DRF using our designed generative networks and loss function named Retinex loss. The major contributions of this work can be summarized as following.
We propose Deep Retinex Fusion (DRF) for adaptive IR and VIS image super-resolution fusion which can obtain excellent performance without any dataset. DRF captures component priors which distangled by physical Retinex model using our designed networks, and then these priors are reconstructed to obtain high-quality super-resolution fusion result using our design loss functions.

- We design a novel dual-path feature switching and skipping network named ZipperNet, which is a generative feature fusion network based on autoencoder. ZipperNet can fuse deep features with delay features switching and obtain high generative quality with skipping connections. Meanwhile, based on Retinex theory, we design novel adaptive fusion loss functions which can balance dynamic range compression, edge enhancement and lightness constancy.

- We implement both IR and VIS image fusion and super-resolution using our unified, dataset-free and self-supervised DRF. Different from methods jointly implementing image fusion and super-resolution using different models, we embed the down-sampling layer behind ZipperNet in DRF. Therefore, exploiting low-level image statistics prior, we can obtain super-resolution fused image only from single pair inputs without any external training data.

This paper is organized as follows. Section II reviews some related works and describes the idea of self-supervised disentangled learning. Section III introduces the design of DRF, the architecture of networks and the design of loss functions. In Section IV, our proposed DRF are compared with 6 state-of-the-art methods on public datasets. In Section V, we discuss the effectiveness of our designed DRF. Finally, we conclude our study in Section VI.

II. RELATED WORKS

In this section, we first review the deep learning-based IR and VIS image fusion methods, and next review the generative learning works and demonstrate the disentangled generative learning.

A. Deep Learning-Based Fusion Methods

The past years have seen a significant development in the study of image fusion using deep learning especially for IR and VIS image fusion. In 2016, Liu et al. [9] applied convolutional neural network (CNN) to combine generative activity level measurement and fusion rule for image fusion according to the training from image patches and handcrafted blurred patches. In 2018, Li et al. [18] applied auto-coder architecture and exploited dense block into encoder part to improve the performance of IR and VIS image fusion. In the same year, Ma et al. [10] first studied IR and VIS image fusion with generative adversarial network (GAN), which generates clean and clear fused images by sharpening IR images with abundant texture details and clear high-lighted targets. In 2020, Zhang et al. [11] proposed a CNN based general image fusion framework IFCNN, which exploits two convolutional layers to extract deep features from input images, fuses them using an appropriate fusion rule, and reconstructs fused image via feature reconstruction. Ma et al. [15] proposed a dual-discriminator conditional GAN for multi-resolution image fusion: the generator of DDcGAN generates fused image based on a content loss to fool the two discriminators, and the two discriminators distinguish the differences between the fused image and inputs (IR and VIS source images), respectively. Xu et al. [19] proposed a unified unsupervised image fusion method named U2Fusion, which estimates the importance of source inputs and proposes adaptive information preservation degrees; therefore, this work can deal with different fusion tasks in high performance with the same sets of parameters. Similar to Unet [20] and Unet++ [21], Li et al. upgraded DenseFuse into NestFuse [22], which first designs a nest connection-based network to extract multiscale deep features, next fuses them in the same scale, and finally reconstructs fused image. In 2021, Li et al. further upgraded NestFuse into RFN-Fuse [18] as a residual fusion network, which proposes a novel detail-preserving loss function and a feature enhancing loss function to improve the fusion performance. Recently, Li et al. [23] integrated multi-scale attention mechanism into both generator and discriminator of GAN (AttentionFGAN) to fuse infrared and visible images.

B. Generative Non-data-driven Learning

Since 2018, DIP [25] was first proposed, in which the generator network structure is sufficient to capture low-level image statistics prior to implement the inverse problems such as super-resolution, denoising and inpainting without any training data. This work provides a new point of view when dealing with inverse problems, e.g. super-resolution can be obtained using only self-similarity of low-resolution images and generator network. Based on DIP, Double-DIP [26] was designed, therefore, the unsupervised layer decomposition tasks of a single image such as segmentation,
watermark-removal and transparency separation can be implemented by coupled DIP networks. Furthermore, Wu et al. [27] proposed an unsupervised method to generate 3D deformable object only using raw single-view images. Its core idea is to disentangle each input image into depth, albedo, viewpoint and illumination, and generate them through end-to-end learning. Significantly, all these works [25], [26], [27] are based on auto-coder (encoder-decoder) architecture, demonstrating that the auto-coder architectures perform well in generative tasks.

However, inspired by above works, we propose a universal generative non-data-driven unsupervised learning method named Deep Retinex Fusion. The core idea of DRF is to disentangle the input(s) into components which are the essential parts for specific physical model (for the task to be solved) via generative networks, and then extract the image low-level statistic priors using network structure. Moreover, in order to implement non-data-driven unsupervised learning, the loss functions are designed based on the specific physical model. Therefore, these loss functions can combine the component priors into a reconstruction closed loop. This process is actually mapping a specific physical model into a deep learning version via networks by exploiting the self-similarity of source input(s), low-level statistics prior of source input(s) and handcrafted prior of networks in order to achieve good performance in super-resolution and fusion.

III. METHOD
In this Method section, we provide the problem formulation, the network architectures and the loss functions for IR and VIS image super-resolution fusion.

A. Deep Retinex Fusion

Based on Retinex theory, the real world is colorless, and the color perceived is the result of interaction between light and object. The basic theory of Retinex is that the object color is determined by the object reflection response to light spectra, rather than by the absolute value of the reflected light intensity. The lightness of the object is consistent and not affected by the illumination non-uniformity. Different from conventional linear and nonlinear methods often used for single-type image feature enhancement, Retinex satisfying lightness consistency can balance dynamic range compression, edge enhancement and color constancy, thus supporting adaptive enhancement of various types of image features. Therefore, the adaptive enhancement capability of Retinex can significantly improve the inconsistency of IR and VIS fusion. In this work, the Retinex theory is exploited as the recorded image \( I(x,y) \) that can be disentangled into reflection image \( R(x,y) \) and lighting image \( L(x,y) \) also described by (1), and the super-resolution fused image of IR and VIS can be treated as the reflection image \( R(x,y) \). Therefore, we formulate the IR and VIS image fusion model as a conditional generative model by ZipperNet, LightingNet and AdjustingNet. As shown in Fig. 2, ZipperNet is designed for generating the super-resolution fused image (reflection image \( R(x,y) \)) and LightingNet is designed for generating the transform maps (lighting image \( L(x,y) \)). Meanwhile, the self-similarity of image and the deep priors are exploited to improve the resolution of the fused image by embedding a down-sampling layer behind the ZipperNet. Additionally, in order to adjust the overall lightness of the generated super-resolution fused image, the AdjustingNet is designed to regress two parameters \( \alpha_1 \) and \( \alpha_2 \). These generative networks are the “CPUs” of the DRF model. Furthermore, the Retinex loss is designed to construct the relation between the input images \( I(x,y) \), the super-resolution fused image \( R(x,y) \), and the transform maps \( L(x,y) \) as well as the adjusting parameters \( \alpha_1 \) and \( \alpha_2 \), therefore, it is the “bus” of the DRF model.

Significantly, the inputs of data-driven methods are different samples in dataset. However, our designed DRF is non-data-driven method, thus its inputs are the fixed sample in epochs.

B. Network Architectures

Section A describes the framework of our designed DRF. ZipperNet which is a dual-path feature switching and skipping network is employed to generate the super-resolution fused image. LightingNets which are the single-path feature skipping networks are employed to generate the transform maps of IR and VIS. AdjustingNet is the same single-path feature skipping network as LightingNets, but only the central (one or two) pixel(s) is/are used as the overall lightness adjusting parameter(s). Essentially, all of these networks are encoder-decoder architectures, which perform significantly well in generative tasks especially image-to-image translation [25], [26], [27], [28]. In this part, ZipperNet, LightingNet and AdjustingNet are described.
1) **ZipperNet:** The architecture of ZipperNet is illustrated in Fig. 3(a). ZipperNet can handle various inputs and fuse their deep features. The inputs are the interpolations of N-scale low-resolution IR and VIS images with the same size. The backbone of ZipperNet is composed of 10 encoder-blocks and 5 decoder-blocks, and they are symmetric with the mirror line. Each encoder-block is composed of two 1-stride 3×3 convolution layers with reflection padding to extract feature, a batch normalization layer to prevent gradient explosions and vanishes, a leaky-RELU activating layer, a 1-stride 3×3 convolution layer and a 2-stride 3×3 convolution layer for down-sampling, a batch normalization layer and a leaky-RELU layer, successively. These encoder-blocks construct the dual-path encoder. In order to better fuse the deep features, features in different odd-even layers are added onto the same-depth layer in the other path. Identically, each decoder-block is composed of a ×2 bilinear layer for up-sampling, a batch normalization layer, two 1-stride 3×3 convolution layers with reflection padding, a batch normalization layer, a leaky-RELU layer, two 1-stride 3×3 convolution layers with reflection padding, a batch normalization layer and a leaky-RELU layer, successively. Furthermore, in order to fuse the features between encoder and decoder parts, the skip connections which are symmetric with the mirror line are adopted to concatenate the features. In the end of the ZipperNet, a sigmoid activation function is adopted to format the value range of the output. The numbers of convolution kernels are 128 in each encoder-block, and 128 in each decoder-block.

2) **LightningNet & AdjustingNet:** In fact, LightningNet and AdjustingNet have the same architecture illustrated in Fig. 3(b). LightningNet and AdjustingNet are composed of 5 encoder-blocks, 5 decoder-blocks and skipping connections which are symmetric with the mirror line. LightningNets are exploited to generate the transform maps. Different to exploiting the whole output map of LightningNet, AdjustingNet only exploits the central pixels of the output map as the weighted parameters to adjust the lightness of the generated image. The numbers of convolution kernels of LightningNet and AdjustingNet are 8, 16, 32, 64 and 128 in each encoder-block and 128, 64, 32, 16 and 8 in each decoder-block.

### C. Loss Functions

The loss functions including Retinex loss, joint gradient loss and lock losses are the “bus” which connect various components. Among them, the Retinex loss is the core of the DRF. The original design of Retinex loss is demonstrated in (2), where $H$ and $W$ are the height and width of the image, $R$ is the generated image of ZipperNet, $L^1$ and $L^2$ are the generated images of LightningNet, $I^1$ and $I^2$ are the input IR and VIS images, $i$ and $j$ are the indexes of pixels. L1 norm is adopted here which has been proved having better performance than L2 norm in super-resolution tasks [30].

\[
L^*_\text{Retinex} = \frac{1}{H \cdot W} \sum_{i,j} \left( |R_{i,j} - L^1_{i,j} - T^1_{i,j}| + |R_{i,j} - L^2_{i,j} - T^2_{i,j}| \right)
\]

(2)

However, the random initialization generated results of $R$ and $L$ have high probability existing zero value, and any zero value in them will cause instability. Meanwhile, untreated maps $R$ and $L$ have high lightness dynamic ranges which can be hardly restricted. Therefore, in order to solve these problems, we apply log on (1) to transfer multiplication into sum (see (3)), and reduce the lightness dynamic ranges of $R$ and $L$, respectively.

\[
\log I(x, y) = \log R(x, y) + \log L(x, y)
\]

(3)

The optimized design of Retinex loss (see (4)) depends on (3).

\[
L^\text{Retinex} = \frac{1}{H \cdot W} \sum_{i,j} \left[ |R_{i,j} - \log L^1_{i,j} + c| + |R_{i,j} - \log L^2_{i,j} + c| 
- \log |L^1_{i,j} - c| + \alpha_1 \cdot |L^1_{i,j} + c + |R_{i,j} - \log L^2_{i,j} + c| 
- \log |L^2_{i,j} - c| + \alpha_2 \cdot |L^2_{i,j} + c - R_{i,j} - \log L^1_{i,j} - c| \right]
\]

(4)

In order to avoid zero and negative values in log, we introduce a small bias $c (=10^{-5})$ and use absolute value operation before applying log. Although the optimized Retinex loss can successfully work, because of additional maps L which act on R, the overall fused image is darker than inputs. Therefore, additional $\alpha_1$ and $\alpha_2$ ($0<\alpha_1, \alpha_2<1$) as the weighted learnable parameters generated by AdjustingNet are introduced to improve the visual lightness perception of fused image.

High-quality IR and VIS fusion requires more high-frequency information. Thus, the maximal gradient map between inputs can almost represent the fused gradient map. The joint gradient loss (5) is designed to force the network focus on high-frequency information. $\nabla^2$ is the Laplacian gradient in (5).

\[
L^\text{grad} = \frac{1}{H \cdot W} \sum_{i,j} \left| \nabla^2 R_{i,j} - \max(\nabla^2 I^1_{i,j}, \nabla^2 I^2_{i,j}) \right|
\]

(5)

Furthermore, there is no limitation on transform maps $L$ and lightness weighted parameters $\alpha_1$ and $\alpha_2$ during iterations. Thus, the $L$ and $\alpha$ lock losses (see (5) and (6)) are designed to limit the value range of $L$, $\alpha_1$ and $\alpha_2$.

\[
L^L = \frac{1}{H \cdot W} \sum_{i,j} \left( |L^1_{i,j} - 1| + |L^2_{i,j} - 1| \right)
\]

(6)

\[
L^\alpha = |\alpha_1 - 0.5| + |\alpha_2 - 0.5|
\]

(7)

Based on the Retinex loss, another lightness lock loss (see (7)) is designed to keep the lightness degree of the generated fused image approaching to the inputs, where $\vec{x}$ is the mean pixel value of $x$.

\[
L^\text{mean} = \frac{1}{H \cdot W} \sum_{i,j} \left( \frac{R_{i,j} - T^1_{i,j} + T^2_{i,j}}{2} \right)
\]

(8)

In general, the total loss is shown as (8), and the values of $\lambda_1$, $\lambda_2$, $\lambda_3$, $\lambda_4$ and $\lambda_5$ are 1, 0.2±0.1, 0.25, 0.25 and 1 according to our experience.

\[
L^\text{total} = \lambda_1 \cdot L^\text{Retinex} + \lambda_2 \cdot L^\text{grad} \\
+ \lambda_3 \cdot L^L + \lambda_4 \cdot L^\alpha + \lambda_5 \cdot L^\text{mean}
\]

(9)
IV EXPERIMENTAL RESULTS

A. Qualitative Experimental Results

In order to verify the performance of proposed DRF, we qualitatively compared it with several existing state-of-the-art models including DDcGAN [15], DenseFuse [18], IFCNN [11], RFN-Nest [24], NestFuse [22] and U2Fusion [19] on $\times 2$ scale of blind super-resolution fusion on FLIR, TNO and VIFB datasets. Additionally, we also compared the $\times 4$ scale of blind super-resolution fusion on FLIR. Since these reported works do not contain the super-resolution function [15], [18], [19], [22], [24] and we implemented super-resolution only using a simple down-sampling layer in DRF, therefore, results obtained via [11], [15], [18], [19], [22], [24] were bicubically up-sampled in order to keep the resolution of comparisons identical. The $\times 2$ scale comparisons are listed in Fig. 4, and we randomly chose three results in each dataset. Additionally, the $\times 4$ scale comparisons are listed Fig. 5, but here we chose results of day and night scenes in FLIR. According to the comparisons in Figs. 4 and 5, our DRF have the best comprehensive performance. DRF can adaptively balance IR and VIS information during image fusion. Actually, as high temperature objects should be focused, an ideal IR and VIS image fusion should be the integration of high-temperature targets in IR image and low temperature background in VIS image. According to Figs. 4 and 5, DRF has the best fusion performance and it can well provide distinctions between day and night scenes. Moreover, DRF can preserve abundant visible texture details as well as high-contrast IR information.

B. Quantitative Experimental Results

We further quantitatively compared our designed DRF with DDcGAN [15], DenseFuse [18], IFCNN [11], RFN-Nest [24], NestFuse [22] and U2Fusion [19] on both $\times 2$ and $\times 4$ scales of blind super-resolution fusion on FLIR dataset (Table 1), and additional $\times 2$ scale on TNO and VIFB datasets (Table 2). The evaluation coefficients compose of mean gradient (MG), cross entropy (CEN), edge intensity (EI) and spatial frequency (SF). MG quantifies the high-frequency contents in the fused image, CEN estimates the similarity of image information distribution between source images and fused image, EI reflects the edge intensity calculated by Sobel operator, and SF describes the richness of the texture details. A high-quality fused image should have high MG, 

\cite{1} https://www.flir.com/oem/adas/adas-dataset-form/
\cite{2} https://figshare.com/articles/TNOImageFusionDataset/1008029
\cite{3} https://github.com/xingchenzhang/VIFB
Fig. 5. We compare our approach against multiple state-of-the-art fusion methods (IFCNN [11], DDCGAN [15], DenseFuse [18], U2Fusion [19], NestFuse [22], RFN-Nest [24]) in ×4 condition on FLIR.

Table I

| Training Data | Scale | DDcGAN | DenseFuse | IFCNN | RFN-Nest | NestFuse | U2Fusion | Ours  |
|---------------|-------|--------|-----------|-------|----------|----------|----------|-------|
| MG ↑          | ×2    | 2.8224 | 2.0512    | 3.6517| 1.9033   | 3.0943   | 3.7888   | **4.0045** |
|               | ×4    | 1.4542 | 1.0667    | 1.8914| 0.9801   | 1.5994   | 1.9575   | **2.2802** |
| CEN ↓         | ×2    | 1.0674 | 1.0572    | 1.1840| 1.3543   | 1.1178   | 1.1254   | **0.7810** |
|               | ×4    | 1.0682 | 1.1348    | 1.2379| 1.3555   | 1.1736   | 1.1276   | **0.7457** |
| EI ↑          | ×2    | 29.6982| 21.1764   | 37.6496| 20.3695 | 32.1763 | 38.4379  | **38.9259** |
|               | ×4    | 15.8393| 11.4791   | 20.5459| 10.6633 | 17.3856 | 21.3315  | **23.3445** |
| SF ↑          | ×2    | 6.2729 | 4.3099    | 7.6609| 3.9975   | 6.9372   | 7.7557   | **9.1759** |
|               | ×4    | 3.2317 | 2.2560    | 3.9545| 2.0771   | 3.5764   | 3.9926   | **5.1841** |
EI and SF, but low CEN. According to both Tables 1 and 2, our proposed DRF has the best performance due to the highest statistical MG, EI and SF and the lowest statistical CEN. Additionally, specific evaluation coefficients corresponding to each sample are listed in Fig. 6.

![Fig. 6. Quantitative comparison of our DRF with 6 state-of-the-art methods (IFCNN [11], DDcGAN [15], DenseFuse [18], U2Fusion [19], NestFuse [22], RFN-Nest [24]) and Bicubic SR in ×2 condition on TNO² and VIFB³. Bold mark the 1st best of the performance.](image)

| Dataset | DDcGAN | DenseFuse | IFCNN | RFN-Nest | NestFuse | U2Fusion | Ours |
|---------|--------|-----------|-------|----------|----------|----------|------|
| MG ↑    | TNO    | 3.5773    | 1.8624 | 3.4286   | 1.6621   | 2.6127   | 3.2500| 4.1126 |
|         | VIFB   | 3.6503    | 2.0003 | 3.6443   | 2.0109   | 2.6237   | 3.2992| 4.1345 |
| CEN ↓   | TNO    | 1.3941    | 1.3038 | 1.5360   | 1.5506   | 1.5001   | 1.4970| 1.0108 |
|         | VIFB   | 6.9814    | 4.6489 | 5.2389   | 6.5581   | 7.2613   | 6.6250| 3.6985 |
| EI ↑    | TNO    | 35.0284   | 18.2814| 33.8043  | 17.5379  | 26.0180  | 33.1625| 35.4144 |
|         | VIFB   | 37.5498   | 20.5047| 37.2774  | 25.1663  | 26.9138  | 34.5566| 39.6117 |
| SF ↑    | TNO    | 6.5438    | 3.6571 | 6.7626   | 5.2561   | 5.3974   | 6.1693| 8.2946 |
|         | VIFB   | 8.5957    | 5.0724 | 9.2192   | 4.9092   | 7.1645   | 7.9476| 11.3476 |

**C. Some Implementation Details**

Different from most data-driven methods, we implemented our DRF only with two fixed IR and VIS inputs in each epoch rather than training dataset. Therefore, DRF is designed for single scene fusion (like conventional iterative methods) rather than general scenes. In addition, most dataset images are in gray both in IR and VIS, and lightness is more important than colors. Therefore, in order to unify the process, we transferred all inputs into gray mode in preprocessing. The networks were trained with a learning rate of $1 \times 10^{-3}$, and iterated for 10000 epochs. All of our experiments were implemented on RTX 3060 GPU.

**V Discussion**

Firstly, in order to verify the effect of our proposed loss functions, we implemented visual ablation experiments in which DRF was trained with different loss functions. Especially in Fig. 7, we discussed use of each component in the loss function described in (9). Fig. 7 lists the fusion results corresponding to our designed loss function in Fig. 7(a), loss function without the learnable adjusting parameters $\alpha_1$ or $\alpha_2$ in Fig. 7(b), loss function without the joint gradient loss in Fig. 7(c), and loss function without the lock losses in Fig. 7(d). According to these comparisons, the learnable adjusting parameters $\alpha_1$ and $\alpha_2$ can effectively equalize the histogram (see Fig. 7(a) and (b)); the joint gradient loss described in (5) can retain more high-frequency details (see Fig. 7(a) and (c)); and the lock losses described in (6), (7) and (8) can constrain the learnable adjusting parameters, reflection maps and lightness level (see Fig. 7(a) and (d)). Moreover, we also tested the Retinex loss in dot mode described in (2), and the fusion result in Fig. 7(e) proves that the Retinex loss in log mode described in (4) can adaptively balance the lightness (see Fig. 7(a) and (e)). Additionally, as shown in Fig. 8, with the limitation of lock losses in (7) and (8), the learnable adjusting parameters $\alpha_1$ and $\alpha_2$ can adaptively converge to 0-1 according to the lightness level of source inputs. All these comparisons demonstrate our designed log-mode loss functions considering learnable adjusting parameters, joint gradient loss and lock losses performed well in this DRF task.

Secondly, DRF is resistant to noise. As shown in Fig. 9, we used [31] blindly estimate the noise level of the sample which contains much noise in TNO. Compared to DDcGAN [15], NestFuse [22] and U2Fusion [19] which perform well, our result has the lowest noise level, proving its good noise resistance capability.
Fig. 7. We study the effect of our proposed loss functions.

Thirdly, we verified the adaptive capability of DRF by adjusting levels of IR images. As shown in Fig. 10, the levels value of original IR image is 1.00, and it was varied from 0.01 to 2.00 via manual adjusting. Results obtained via DDcGAN [15], DenseFuse [18], IFCNN [11], RFN-Nest [24], NestFuse [22] and U2Fusion [19] by fusing VIS image and IR image in different levels (2.00, 1.00, 0.50, 0.30, 0.01) are listed in Fig. 10. As the levels of IR image decrease, the IR images gradually darken, and the fusion results obtained by DenseFuse, IFCNN, RFN-Nest and U2Fusion significantly become dark. While as the levels of IR image increase, the image contrast of the IR images gradually reduces, and the fusion result obtained by DenseFuse, IFCNN, RFN-Nest, NestFuse and U2Fusion also demonstrate lower contrast. Though DDcGAN performs satisfied adaptive capability, it fusion quality is still poor. Our proposed DRF always provides high-quality fused images in consist image intensity and contrast even in different level conditions. Therefore, our proposed DRF can adaptively fuse IR and VIS images, thus remarkably improving the fusion robustness.

Fig. 8. We study the convergence of learnable adjusting parameters $\alpha_1$ and $\alpha_2$, black one is $\alpha_1$, and red one is $\alpha_2$.

Fig. 9. Noise resistance capability comparison of our DRF with 3 state-of-the-art methods (DDcGAN [15], U2Fusion [19], NestFuse [22]) on sample in TNO which contains much noise.

VI. CONCLUSION

In this study, we propose DRF as a dataset-free self-supervised disentangled learning method for adaptive IR and VIS image super-resolution fusion. Generative networks ZipperNet, LightingNet and AdjustingNet and Retinex theory based adaptive fusion loss functions are designed for high performance super-resolution image fusion using a unified, dataset-free and self-supervised model. Compared to many state-of-the-art methods, DRF has the best super-resolution fusion performance; and more importantly, DRF can adaptively balance IR and VIS information and has good noise immunity.

ACKNOWLEDGMENT

This work is supported by National Natural Science Foundation of China (61705092) and Natural Science Foundation of Jiangsu Province (BK20200588, BK20180598, BK20170194).

REFERENCES

[1] N. Yamamoto, T. Saito, S. Ogawa, and I. Ishimaru, “Middle infrared (wavelength range: 8 µm-14 µm) 2-dimensional spectroscopy (total weight with electrical controller: 1.7 kg, total cost: less than 10,000 USD) so-called hyperspectral camera for unmanned air vehicles like drones,” Algorithms Technol. Multispectral, Hyperspectral, Ultraspectral Imag. XXII, vol. 9840, p. 984028, 2016, doi: 10.1117/12.2223358.
[2] M. T. Eismann et al., “Comparison of infrared imaging hyperspectral sensors for military target detection applications.” Proceedings of SPIE - The International Society for Optical Engineering vol. 2819, pp. 91–101.
[3] U. Willer, M. Sarajii, A. Khorsandi, P. Geiser, and W. Schade, “Near- and mid-infrared laser monitoring of industrial processes, environment and security applications,” Opt. Lasers Eng., vol. 44, no. 7, pp. 699–710, 2006, doi: 10.1016/j.optlaseng.2005.04.015.
[4] R. Mulaveesala and S. V. Ghali, “Coded excitation for infrared non-destructive testing of carbon fiber reinforced plastics Coded excitation for infrared non-destructive testing of carbon fiber,” vol. 054902, no. 2011, 2013, doi: 10.1063/1.3594551.
[5] Y. Iwasaki, “A method of robust moving vehicle detection for bad weather using an infrared thermography camera,” Proc. 2008 Int. Conf. Wavelet Anal. Pattern Recognition, ICWAPR, vol. 1, pp. 86–90, 2008, doi: 10.1109/ICWAPR.2008.4635755.
Fig. 10. We compare the adaptive capability of our approach against multiple state-of-the-art fusion methods (IFCNN [11], DDcGAN [15], DenseFuse [18], U2Fusion [19], NestFuse [22], RFN-Nest [23]) in ×2 condition on sample in TNO². We leave the VIS image unchanged, adjust the levels of the IR image and obtain the fused results.

[6] J. Phillips, “Evaluation of the fundamental properties of quantum dot infrared detectors,” J. Appl. Phys., vol. 91, no. 7, pp. 4590–4594, 2002, doi: 10.1063/1.1455130.

[7] B. B. Lahiri, S. Bagavathiappan, T. Jayakumar, and J. Philip, “Medical applications of infrared thermography: A review,” Infrared Phys. Technol., vol. 55, no. 4, pp. 221–235, 2012.
J. Ma, H. Xu, J. Jiang
“Probably Symmetric Deformable 3D Objects from Images in the
2019 Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 1, pp. 9446–9454, 2018, doi: 10.1109/CVPR.2018.00984.
IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work., vol. 2020–June, pp. 468–478, 2020, doi: 10.1109/CVPRW.2020.00060.

H. Zhao, O. Gallo, I. Froisio, and J. Kautz, “Loss Functions for Neural Networks for Image Processing,” pp. 1–11, 2015, [Online]. Available: http://arxiv.org/abs/1511.08861.

Yuanjie Gu received the B.E. degree from the School of Petroleum, China University of Petroleum, China, in 2020. He is currently a master student in the Computational Optics Laboratory, School of Science, Jiangnan University, Wuxi. His current research interests include computational photography and deep learning.

Zhibo Xiao received the B.E. degree from the School of Petroleum, China University of Petroleum, China, in 2020. He is currently a master student in the Computational Optics Laboratory, School of Science, Jiangnan University, Wuxi. His current research interests include computational photography and deep learning.

Hailun Wang received the B.E. degree from the School of Petroleum, China University of Petroleum, China, in 2021. She is currently a master student in the Computational Optics Laboratory, School of Science, Jiangnan University, Wuxi. Her current research interests include computational photography and deep learning.

Cheng Liu received the B.E. and M.E. degrees from Soochow University in 1993 and 1996, respectively, and the Ph.D. degree from Shanghai Institute of Optics and Fine Mechanics in 2003. From 2003 to 2011 he was a postdoctoral associate with Yonsei University, National University of Singapore, Guangji Institute of Science and Technology, and University of Sheffield. Since 2011, he becomes a full professor in Jiangnan University and leads the Computational Optics Laboratory. He is also affiliated with Shanghai Institute of Optics and Fine Mechanics. He has published more than 150 journal articles and conference papers. His current research interests include computational imaging, optical testing and microscopy.

Shouya Wang received the B.E. and Ph.D degrees from Nanjing University of Science and Technology in 2010 and 2015. He is currently an associate professor and co-principal investigator in the Computational Optics Laboratory, School of Science, Jiangnan University, Wuxi. He is also affiliated with Single Molecule Nanometry Laboratory in Nanjing Agricultural University. He has published more than 100 journal articles and conference papers. His current research interests include computational sensing and imaging, microscopy and single molecule techniques.