DIDFuse: Deep Image Decomposition for Infrared and Visible Image Fusion

Zixiang Zhao[1] , **Shuang Xu**[1*] , **Chunxia Zhang**[1] , **Junmin Liu**[1] , **Jiangshe Zhang**[1†] and **Pengfei Li**[2]

[1]School of Mathematics and Statistics, Xi'an Jiaotong University, China
[2]Hikvision, China
{zixiangzhao, shuangxu}@stu.xjtu.edu.cn, {cxzhang, junminliu, jszhang}@mail.xjtu.edu.cn,
lipengfei27@hikvision.com

Abstract

Infrared and visible image fusion, a hot topic in the field of image processing, aims at obtaining fused images keeping the advantages of source images. This paper proposes a novel auto-encoder (AE) based fusion network. The core idea is that the encoder decomposes an image into background and detail feature maps with low- and high-frequency information, respectively, and that the decoder recovers the original image. To this end, the loss function makes the background/detail feature maps of source images similar/dissimilar. In the test phase, background and detail feature maps are respectively merged via a fusion module, and the fused image is recovered by the decoder. Qualitative and quantitative results illustrate that our method can generate fusion images containing highlighted targets and abundant detail texture information with strong reproducibility and meanwhile surpass state-of-the-art (SOTA) approaches.

1 Introduction

Image fusion is an image processing technique for information enhancement. The principle is to preserve the complementary and redundant information from source images containing the same scene without artifacts [Meher *et al.*, 2019]. In image fusion, the infrared and visible image fusion, a.k.a. IVIF, can be applied to many domains, such as surveillance [Bhatnagar and Liu, 2015], modern military and fire rescue tasks [Lahoud and Susstrunk, 2018; Hu *et al.*, 2017], face recognition [Ma *et al.*, 2016] and so on.

Proverbially, infrared images can avoid visually cognitive obstacles caused by illumination changes and artifacts, but they are with low spatial resolution and poor texture detail information. Conversely, visible images are with high spatial resolution and rich information of appearance and gradient, while they are easily affected by obstructions and light reflections. Therefore, making the fusion image retain both thermal radiation information of the infrared images and gradient information of the visible images will be conducive to target recognition and tracking.

In general, the IVIF algorithms can be divided into two groups: traditional methods and deep learning methods. Specifically, representative traditional methods include image multi-scale transformation [Li *et al.*, 2011], sparse representation [Zong and Qiu, 2017], subspace learning [Patil and Mudengudi, 2011] and the saliency based method [Zhang *et al.*, 2017].

Currently, deep learning (DL) has emerged as a prevalent tool in the field of IVIF. DL based methods can be categorized into three groups. The first group is based on Generative Adversarial Networks (GANs). In FusionGAN [Ma *et al.*, 2019b], a generator creates fused images with infrared thermal radiation and visible gradient information and a discriminator forces the fused images to have more details from the visible images. In the light of Conditional GANs [Mirza and Osindero, 2014], detail preserving GAN [Ma *et al.*, 2020] changes the loss function of FusionGAN for improving the quality of detail information and sharpening the target boundary. The second group [Li *et al.*, 2018; Lahoud and Süsstrunk, 2019] is an extension of image multi-scale transformation. Generally speaking, they transform images from the spatial domain to background and detail domains by means of filters or optimization based methods. Background images are simply averaged. Since there are high-frequency textures in detail images, they fuse feature maps of detail images extracted from a pre-trained network (for example, VGG [Simonyan and Zisserman, 2014]). At last, a fusion image is recovered by merging the fused background and detail images. The third group consists of AE based methods [Li and Wu, 2018]. In the training phase, an AE network is trained. In the test phase, they fuse feature maps of source images, which then pass through the decoder to recover a fusion image. In summary, in DL based methods, deep neural networks (DNNs) are often employed to extract features of input images and then a certain fusion strategy is exploited to combine features to complete the image fusion task.

It is worth pointing out a shortcoming of the second group, i.e. DL is used only in the fusion stage and they employ filters or optimization based methods in the decomposition stage. To overcome this shortcoming, by combining principles of the second and third groups, we propose a novel IVIF network,

---

*Co-first Author
†Corresponding Author

called deep image decomposition based IVIF (DIDFuse). Our contributions are two-fold:

(1) To the best of our knowledge, this is the first deep image decomposition model for IVIF task, where both fusion and decomposition are accomplished via an AE network. The encoder and the decoder are responsible for image decomposition and reconstruction, respectively. In training phase, for decomposition stage, the loss function forces background and detail feature maps of two source images similar/dissimilar. Simultaneously, for reconstruction stage, the loss function maintains pixel intensities between source and reconstructed images, and gradient details of the visible image. In the test phase, background and detail feature maps of test pairs are separately fused according to a specific fusion strategy, and then the fused image can be acquired through the decoder.

(2) As far as we know, the performance of existing IVIF methods [Ma *et al.*, 2016; Li and Wu, 2018; Zhang *et al.*, 2017; Li *et al.*, 2018] is only verified on a limited number of hand-picked examples in TNO dataset. However, we test our model on three datasets, including TNO, FLIR and NIR. In total, there are 132 test images with indoor and outdoor scenes, and with daylight and nightlight illuminations. Compared with SOTA methods, our method can robustly create fusion images with brighter targets and richer details. It can be potentially utilized in target recognition and tracking.

The remaining article is arranged as follows. Related work is introduced in section 2. The mechanism of the proposed network is described in section 3. Then, experimental results are reported in section 4. At last, some conclusions are drawn in section 5.

2 Related Work

Since our network structure is closely related with U-Net, we introduce U-Net in section 2.1. Then, traditional two-scale image decomposition methods are briefly reviewed in section 2.2.

2.1 U-Net and Skip Connection

U-Net is applied to biomedical image segmentation[Ronneberger *et al.*, 2015], similar to AE network, U-Net consists of a contracting path for feature extraction and an expanding path for precise localization. Compared with AE, there is a channel-wise concatenation of corresponding feature maps from contracting and expanding paths in U-Net. In this manner, it can extract "thicker" features that help preserve image texture details during downsampling. In literature [Mao *et al.*, 2016], a U-Net-like symmetric network is used for image restoration. It employs skip connection technique, where feature maps of convolution layers are added to corresponding deconvolution layers to enhance the information extraction capability of the neural network and to accelerate convergence.

2.2 Two-Scale Decomposition

As a subset of multi-scale transformation, two-scale decomposition in IVIF decomposes an original image into background and detail images with background and target information, respectively. In [Li and Wu, 2018], given an image I, they

obtained the background image I^b by solving the following optimization problem,

$$I^b = \arg\min ||I - I^b||_F^2 + \lambda(||g_x * I^b||_F^2 + ||g_y * I^b||_F^2),$$

where $*$ denotes a convolution operator, and $g_x = [-1, 1]$ and $g_y = [-1, 1]^T$ are gradient kernels. Then, the detail image is acquired by $I^d = I - I^b$. Similarly, a box filter is used to get the background image in [Lahoud and Süsstrunk, 2019], and the method of obtaining the detail image is the same as that of [Li and Wu, 2018]. After decomposition, background and detail images are separately fused with different criteria. At last, the fused image is reconstructed by combining fused background and detail images.

3 Method

In this section, we will introduce our DIDFuse algorithm and the proposed network structure. In addition, details of training and testing phases are also illustrated.

3.1 Motivation

As described in section 2.2, two-scale decomposition decomposes the input image into a background image containing low-frequency information with large-scale pixel intensity changes and a detail image embodying high-frequency information with small-scale pixel intensity changes. Currently, most algorithms incorporate certain prior knowledge, and employ filters or optimization based methods to decompose images. Hence, they are manually designed decomposition algorithms. We highlight that image decomposition algorithms are intrinsically feature extractors. Formally, they transform source images from spatial domain into feature domain. It is well known that the DNN is a promising data-driven feature extractor and has great superiority over traditional manually-designed methods. Unfortunately, it lacks a DL based image decomposition algorithm for IVIF task.

Consequently, we present a novel deep image decomposition network in which an encoder is exploited to perform two-scale decomposition and extract different types of information, and a decoder is used to recover original images.

3.2 Network Architecture

Our neural network consists of an encoder and a decoder. As illustrated in Figure 1, the encoder is fed with an infrared or a visible image and generates background and detail feature maps. Then, the network concatenates two kinds of feature maps along channels. At last, concatenated feature maps pass through decoder to recover the original image. To prevent the detail information of the feature maps from being lost after multiple convolutions and to speed up the convergence, we add the feature maps from the first and second convolutions to the inputs of the last and penultimate convolutions, and the adding strategy is concatenating the corresponding feature maps along channels. As a consequence, the pixel intensity and gradient information of the source images can be better retained in the reconstructed image.

Table 1 lists the network configuration. Encoder and decoder contain four and three convolutional layers, respectively. Each layer consists of a padding, a 3×3 convolution, a batch
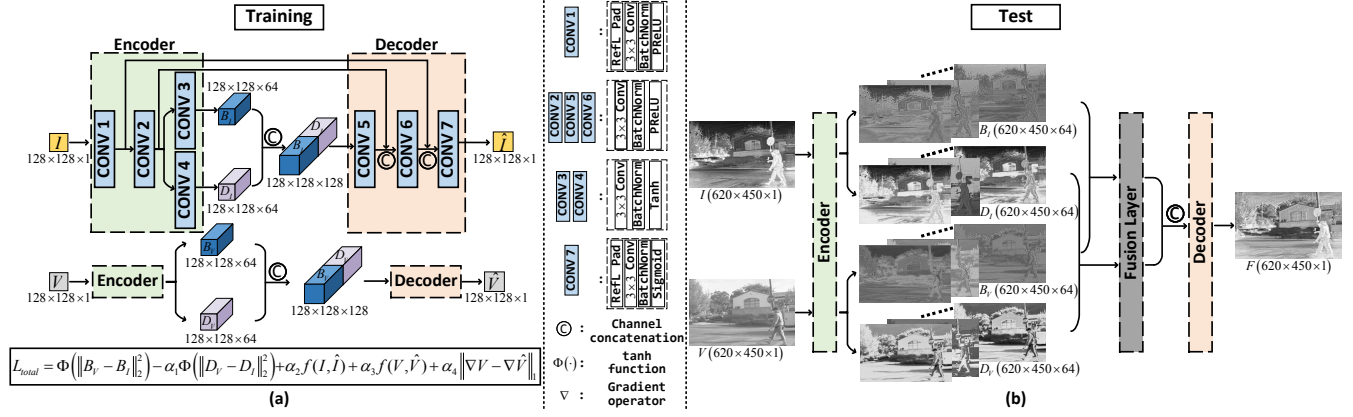
Figure 1: Neural network framework of DIDFuse.

Layers	Size	InC	OutC	Padding	Activation
conv1	3	1	64	Reflection	PReLU
conv2	3	64	64	Zero	PReLU
conv3	3	64	64	Zero	Tanh
conv4	3	64	64	Zero	Tanh
conv5	3	128	64	Zero	PReLU
conv6	3	64	64	Zero	PReLU
conv7	3	64	1	Reflection	Sigmoid

Table 1: Network configuration. Size denotes the size of convolutional kernel. InC and OutC are the numbers of input and output channels, respectively.

normalization and an activation function. The first and the last layers utilize reflection padding to prevent artifacts at the edges of the fused image. Activation functions of `conv3` and `conv4` are set to the hyperbolic tangent function (tanh) since they output background and detail feature maps. As for `conv7`, it is activated by sigmoid function since it reconstructs original images. Other layers are followed by parametric rectified linear units (PReLU).

3.3 Loss Function

In the training phase, we aim to obtain an encoder that performs two-scale decomposition on the source images, and at the same time, acquire a decoder that can fuse the images and preserve the information of source images well. The training process is shown in Figure 1(a).

Image decomposition. Background feature maps are used to extract the common features of source images, while detail feature maps are used to capture the distinct characteristics from infrared and visible images. Therefore, we should make the gap of background feature maps small. In contrast, the gap of detail feature maps should be great. To this end, the loss function of image decomposition is defined as follow,

$$L_1 = \Phi\left(\|B_V - B_I\|_2^2\right) - \alpha_1 \Phi\left(\|D_V - D_I\|_2^2\right), \quad (1)$$

where B_V, D_V are the background and detail feature maps of the visible image V, and B_I, D_I are those of the infrared

image I. $\Phi(\cdot)$ is tanh function that is used to bound gap into interval $(-1, 1)$.

Image Reconstruction

As for image reconstruction, to successfully retain the pixel intensity and detailed texture information of input images, the reconstruction loss function is given by

$$L_2 = \alpha_2 f(I, \hat{I}) + \alpha_3 f(V, \hat{V}) + \alpha_4 \left\|\nabla V - \nabla \hat{V}\right\|_1, \quad (2)$$

where I and $\hat{I}$, V and $\hat{V}$ represent the input and reconstructed images of infrared and visible images, respectively. ∇ denotes the gradient operator, and

$$f(X, \hat{X}) = \left\|X - \hat{X}\right\|_2^2 + \lambda L_{SSIM}(X, \hat{X}), \quad (3)$$

where X and $\hat{X}$ represent the above input image and the reconstructed image, and λ is the hyperparameter. SSIM is the structural similarity index[Wang *et al.*, 2004], which is a measure of the similarity between two pictures. Then L_{SSIM} can be described as

$$L_{SSIM}(X, \hat{X}) = \frac{1 - SSIM(X, \hat{X})}{2}.$$

Remark that L_2-norm measures the pixel intensity agreement between original and reconstructed images, and that L_{SSIM} computes image dissimilarity in terms of brightness, contrast and structure. Specially, since visible images are with enriched textures, the reconstruction of visible images is regularized by gradient sparsity penalty to guarantee texture agreement.

Combining Eqs. (1) and (2), the total loss L_{total} can be expressed as

$$\begin{aligned} L_{total} =& L_1 + L_2 \\ =& \Phi\left(\|B_V - B_I\|_2^2\right) - \alpha_1 \Phi\left(\|D_V - D_I\|_2^2\right) \\ &+ \alpha_2 f(I, \hat{I}) + \alpha_3 f(V, \hat{V}) + \alpha_4 \left\|\nabla V - \nabla \hat{V}\right\|_1, \end{aligned} \quad (4)$$

where $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ are the tuning parameters.

3.4 Fusion Strategy

In the above subsections, we have proposed network structure and loss function. After training, we will acquire a decomposer (or say, encoder) and a decoder. In the test phase, we aim to fuse infrared and visible images. The workflow is shown in Figure1(b). Different from training, a fusion layer is inserted in the test phase. It fuses background and detail feature maps separately. In formula, there is

$$B_F = \text{Fusion}(B_I, B_V), D_F = \text{Fusion}(D_I, D_V), \quad (5)$$

where B_F and D_F denote the fused background and detail feature maps, respectively. In this paper, three fusion strategies are considered as follows:

- Summation method: $B_F = B_I \oplus B_V, D_F = D_I \oplus D_V$, where the symbol $\oplus$ means element-wise addition.

- Weighted average method: $B_F = \gamma_1 B_I \oplus \gamma_2 B_V, D_F = \gamma_3 D_I \oplus \gamma_4 D_V$, where $\gamma_1 + \gamma_2 = \gamma_3 + \gamma_4 = 1$ and the default settings for $\gamma_i (i = 1, \cdots, 4)$ are all equal to 0.5.

- L_1-norm method: Referring to [Li and Wu, 2018], we use the L_1-norm as a measure of activity, combining with the softmax operator. In detail, we can obtain the activity level map of the fused background and detail feature maps by $\|B_i(x,y)\|_1$ and $\|D_i(x,y)\|_1 (i = 1, 2)$, where B_1, B_2, D_1 and D_2 represent B_I, B_V, D_I and D_V, and (x, y) represents the corresponding coordinates of the feature maps and the fused feature map. Then the adding weights can be calculated by:

$$\eta_i^B(x,y) = \frac{\psi\left(\|B_i(x,y)\|_1\right)}{\sum_{i=1}^2 \psi\left(\|B_i(x,y)\|_1\right)},$$
$$\eta_i^D(x,y) = \frac{\psi\left(\|D_i(x,y)\|_1\right)}{\sum_{i=1}^2 \psi\left(\|D_i(x,y)\|_1\right)},$$

where $\psi(\cdot)$ is a 3×3 box blur (also known as a mean filter operator). Consequently, we have

$$B_F = (\eta_1^B \otimes B_I) \oplus (\eta_2^B \otimes B_V),$$
$$D_F = (\eta_1^D \otimes D_I) \oplus (\eta_2^D \otimes D_V).$$

where $\otimes$ means element-wise multiplication.

4 Experiment

The aim of this section is to study the performance of our proposed model and compare it with other SOTA models, including FusionGAN [Ma et al., 2019b], Densefuse [Li and Wu, 2018], ImageFuse [Li et al., 2018], DeepFuse [Prabhakar et al., 2017], TSIFVS [Bavirisetti and Dhuli, 2016], TVADMM [Guo et al., 2017], CSR [Liu et al., 2016] and ADF [Bavirisetti and Dhuli, 2015]. All experiments were conducted with Pytorch on a computer with Intel Core i7-9750H CPU@2.60GHz and RTX2070 GPU.

We employ six metrics to evaluate the quality of a fused image, that is, entropy (EN), mutual information (MI), standard deviation (SD), spatial frequency (SF), visual information fidelity (VIF) and average gradient (AG). More details for these metrics can be seen in [Ma et al., 2019a].

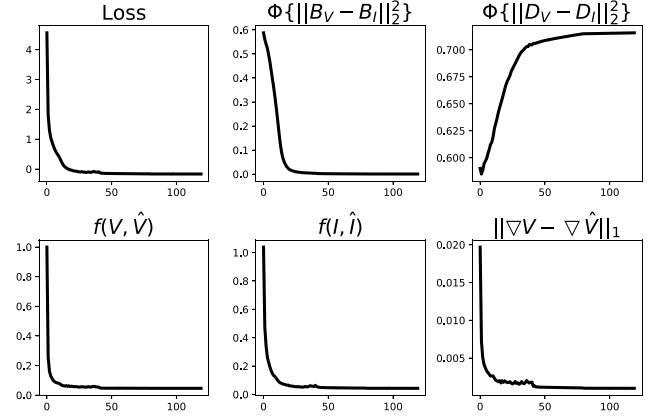|	Dataset(pairs)	Illumination
Training	FLIR-Train(180)	Daylight&Nightlight
Validation	NIR-Urban(58)	Daylight
	NIR-Street(50)	Daylight
Test	TNO (40)	Nightlight
	FLIR-Test(40)	Daylight&Nightlight
	NIR-Country(52)	Daylight

Table 2: Dataset used in this paper.

Figure 2: Loss curves over 120 epochs.

Datasets and preprocessing. Our experiments are conducted on three datasets, including TNO [Toet and Hogervorst, 2012], NIR [Brown and Süsstrunk, 2011] and FLIR (available at https://github.com/jiayi-ma/RoadScene). In our experiment, we divide them into training, validation, and test sets. Table 2 shows the numbers of image pairs, illumination and scene information of the datasets. We randomly selected 180 pairs of images in the FLIR dataset as training samples. Before training, all images are transformed into grayscale. At the same time, we center-crop them with 128×128 pixels.

Hyperparameters setting. The tuning parameters in loss function are empirically set as follows: $\alpha_1 = 0.05$, $\alpha_2 = 2$, $\alpha_3 = 2$, $\alpha_4 = 10$ and $\lambda = 5$. In training phase, the network is optimized by Adam over 120 epochs with a batch size of 24. As for learning rate, we set it to 10^{-3} and decrease it by 10 times every 40 epochs. Figure 2 displays loss curves versus epoch index. It is shown that all loss curves are very flat after 120 epochs. In other words, the network is able to converge with this configuration.

4.1 Experiments on Fusion Strategy

As described in section 3.4, fusion strategy plays an important role in our model. We investigate the performance of three strategies on validation set. Table 3 reports numerical results of six metrics on validation set. Obviously, it is shown that summation strategy achieves higher values, especially in terms of SD, SF, VIF and AG. Hence, the following experiments adopt summation strategy.

Dataset: NIR Dataset. Scene: Street			
Method	Summation	Average	L_1-norm
EN	**7.17 ± 0.10**	6.85 ± 0.03	6.87 ± 0.03
MI	**4.69 ± 0.06**	4.68 ± 0.04	4.68 ± 0.03
SD	**55.51 ± 1.74**	36.51 ± 0.59	36.88 ± 0.65
SF	**24.28 ± 1.12**	16.85 ± 0.23	16.80 ± 0.23
VIF	**1.02 ± 0.04**	0.62 ± 0.01	0.63 ± 0.01
AG	**7.18 ± 0.41**	4.89 ± 0.07	4.88 ± 0.07

Dataset: NIR Dataset. Scene: Urban			
Method	Summation	Average	L_1-norm
EN	**7.18 ± 0.14**	7.12 ± 0.03	7.12 ± 0.03
MI	6.07 ± 0.07	**6.15 ± 0.04**	6.14 ± 0.03
SD	**61.46 ± 1.55**	41.64 ± 0.51	41.73 ± 0.50
SF	**29.22 ± 1.14**	20.20 ± 0.22	20.17 ± 0.22
VIF	**1.13 ± 0.05**	0.77 ± 0.01	0.77 ± 0.01
AG	**8.13 ± 0.41**	5.85 ± 0.06	5.85 ± 0.06

Table 3: Results of validation set for choosing the addition strategy.

4.2 Experiments on Image Decomposition

One of our contributions is the deep image decomposition. It is interesting to study whether decomposed feature maps are able to meet our demands. In Figure 3, it displays the first channels of feature maps which are generated by `conv3` and `conv4`. It is evident that our method can separate the backgrounds and details of infrared and visible images. For background feature maps, it is found that B_I and B_V are visually similar, and they reflect the background and environment of the same scene. Conversely, the gap between D_I and D_V is large, which illustrates the distinct characteristics contained in different source images. That is, the infrared images contain target highlight and thermal radiation information while gradient and texture information of targets are involved in the visible images. In conclusion, it to some degree verifies the rationality of our proposed network structure and image decomposition loss function.

4.3 Comparison with Other Models

In this subsection, we will compare our model with the other popular counterparts.

Qualitative comparison. Figure 4 exhibits several representative fusion images generated by different models. Visual inspection shows that, in the images containing people, other methods have problems such as weak high-lighted objects, poor contrast and less prominent contour of targets and backgrounds. Similarly, if the images are natural landscapes, others have blurred boundaries of mountains and trees, poor color contrast, and insufficient sharpness. Conversely, our method can obtain fused images with brighter targets, sharper edge contours and retaining richer detailed information.

Quantitative comparison. Subsequently, quantitative comparison results on test set are listed in Table 4. It is found that our model is the best performer on all datasets in terms of all metrics. As for competitors, they may perform well on a dataset in terms of part of metrics. This result demonstrates
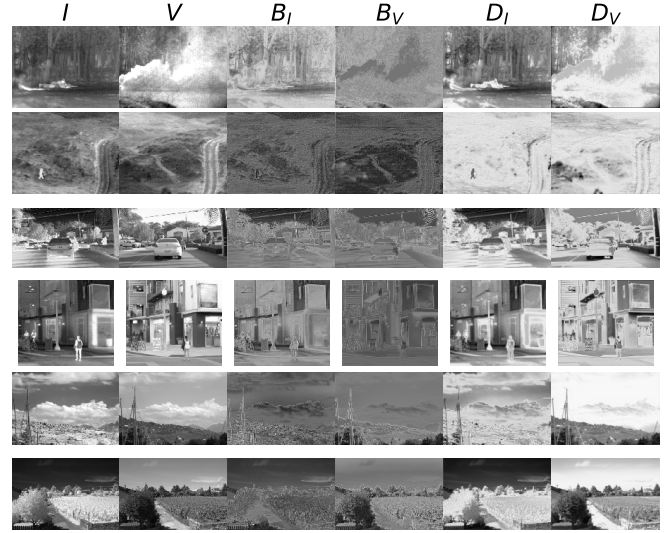

Figure 3: Illustration of deep image decomposition. From left to right: infrared image, visible image, background and detail feature maps of infrared image and visible image.
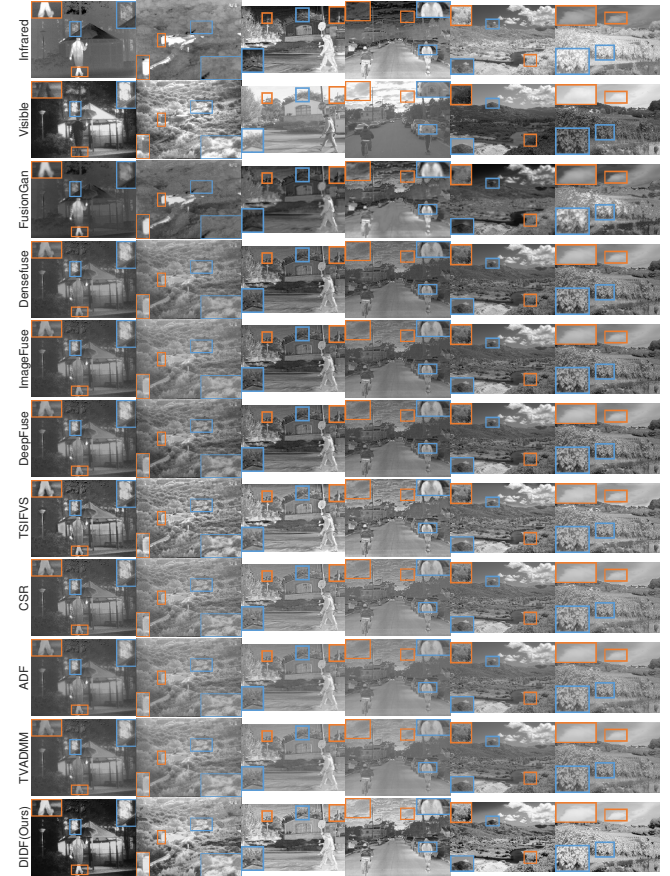

Figure 4: Qualitative results for different methods. Areas marked by orange and blue boxes are amplified for ease of inspection.

that images fused by our model are with enriched textures and

Dataset: TNO image fusion dataset

Metrics	FusionGAN	DenseFuse	ImageFuse	DeepFuse	TSIFVS	TVADMM	CSR	ADF	DIDFuse
EN	6.576	6.842	6.382	6.860	6.669	6.402	6.428	6.399	**7.006**
MI	2.341	2.302	2.155	2.298	1.717	2.041	1.990	2.007	**2.347**
SD	29.035	31.817	22.938	32.249	28.036	23.007	23.603	22.963	**42.657**
SF	8.762	11.095	9.800	11.125	12.598	9.034	11.445	10.782	**13.126**
VIF	0.258	0.572	0.306	0.581	0.456	0.284	0.312	0.286	**0.623**
AG	2.417	3.597	2.719	3.599	3.980	2.518	3.367	2.988	**4.294**

Dataset: FLIR image fusion dataset

Metrics	FusionGAN	DenseFuse	ImageFuse	DeepFuse	TSIFVS	TVADMM	CSR	ADF	DIDFuse
EN	7.017	7.213	6.992	7.213	7.152	6.797	6.909	6.798	**7.344**
MI	2.684	2.727	2.783	2.725	2.312	2.473	2.569	2.721	**2.882**
SD	34.383	37.315	32.579	37.351	35.889	28.071	30.529	28.371	**46.888**
SF	11.507	15.496	14.519	15.471	18.794	14.044	17.128	14.480	**18.835**
VIF	0.289	0.498	0.419	0.498	0.503	0.325	0.373	0.337	**0.538**
AG	3.205	4.822	4.150	4.802	5.568	3.524	4.799	3.564	**5.574**

Dataset: RGB-NIR Scene Dataset

Metrics	FusionGAN	DenseFuse	ImageFuse	DeepFuse	TSIFVS	TVADMM	CSR	ADF	DIDFuse
EN	7.055	7.304	7.217	7.303	7.300	7.129	7.170	7.105	**7.357**
MI	3.003	**4.045**	3.967	4.040	3.285	3.673	3.699	3.944	3.795
SD	34.912	45.850	42.307	45.815	43.743	40.469	40.383	38.978	**57.557**
SF	14.309	18.718	18.360	18.627	20.646	16.685	20.370	17.313	**26.529**
VIF	0.424	0.677	0.613	0.676	0.688	0.530	0.583	0.538	**0.937**
AG	4.564	6.228	5.920	6.178	6.823	5.319	6.488	5.381	**8.741**

Table 4: Quantitative results of different methods. The largest value is shown in bold, and the second largest value is underlined.

satisfy human visual system.

4.4 Experiments on Reproducibility

As is known, deep learning methods are often criticized for instability. Therefore, we test the reproducibility of DIDFuse in the last experiment. We repeatedly train the network 25 times and quantitatively compare the 25 parallel results. As shown in Figure 5, the black solid curves report six metrics over 25 experiments. The red dashed line and blue dotted line represent the greatest and the second greatest values in the comparison methods, respectively. Similar to the above results, our method can basically keep the first place all the time, indicating that DIDFuse can generate high-quality fused images steadily.
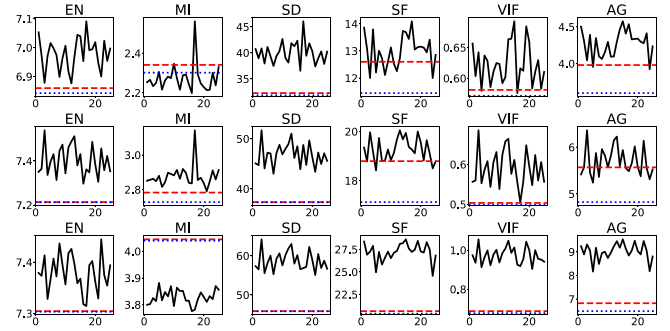

Figure 5: Test results of model reproducibility. From top to bottom: image fusion dataset TNO, FLIR, and NIR. From left to right: the values of EN, MI, SD, SF, VIF and AG.

5 Conclusion

To solve the IVIF issue, we construct a new AE network in which the encoder is used for two-scale image decomposition and the decoder is responsible for image reconstruction. In the training phase, the encoder is trained to output background and feature maps, then the decoder reconstructs original images. In the test phase, we set a fusion layer between the encoder and decoder to fuse background and detail feature maps through a specific fusion strategy. Finally, the fused image can be acquired through the decoder. We test our model on TNO, FLIR, and NIR datasets. Qualitative and quantitative results show that our model outperforms other SOTA methods, since

our model can steadily obtain a fusion image of highlighted targets and rich details.

References

[Bavirisetti and Dhuli, 2015] Durga Prasad Bavirisetti and Ravindra Dhuli. Fusion of infrared and visible sensor images based on anisotropic diffusion and karhunen-loeve transform. *IEEE Sensors Journal*, 16(1):203–209, 2015.

[Bavirisetti and Dhuli, 2016] Durga Prasad Bavirisetti and Ravindra Dhuli. Two-scale image fusion of visible and infrared images using saliency detection. *Infrared Physics & Technology*, 76:52–64, 2016.

[Bhatnagar and Liu, 2015] Gaurav Bhatnagar and Zheng Liu. A novel image fusion framework for night-vision navigation and surveillance. *Signal, Image and Video Processing*, 9(1):165–175, 2015.

[Brown and Süsstrunk, 2011] Matthew Brown and Sabine Süsstrunk. Multi-spectral sift for scene category recognition. In *CVPR 2011*, pages 177–184. IEEE, 2011.

[Guo et al., 2017] Hanqi Guo, Yong Ma, Xiaoguang Mei, and Jiayi Ma. Infrared and visible image fusion based on total variation and augmented lagrangian. *Journal of the Optical Society of America A*, 34(11):1961–1968, 2017.

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

[Lahoud and Susstrunk, 2018] F. Lahoud and S. Susstrunk. Ar in vr: Simulating infrared augmented vision. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3893–3897, Oct 2018.

[Lahoud and Süsstrunk, 2019] Fayez Lahoud and Sabine Süsstrunk. Fast and efficient zero-learning image fusion. *arXiv preprint arXiv:1905.03590*, 2019.

[Li and Wu, 2018] Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2018.

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

[Li et al., 2018] Hui Li, Xiao-Jun Wu, and Josef Kittler. Infrared and visible image fusion using a deep learning framework. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2705–2710. IEEE, 2018.

[Liu et al., 2016] Yu Liu, Xun Chen, Rabab K Ward, and Z Jane Wang. Image fusion with convolutional sparse representation. *IEEE Signal Processing Letters*, 23(12):1882–1886, 2016.

[Ma et al., 2016] Jiayi Ma, Chen Chen, Chang Li, and Jun Huang. Infrared and visible image fusion via gradient transfer and total variation minimization. *Information Fusion*, 31:100–109, 2016.

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

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

[Ma et al., 2020] Jiayi Ma, Pengwei Liang, Wei Yu, Chen Chen, Xiaojie Guo, Jia Wu, and Junjun Jiang. Infrared and visible image fusion via detail preserving adversarial learning. *Information Fusion*, 54:85–98, 2020.

[Mao et al., 2016] Xiaojiao Mao, Chunhua Shen, and Yu-Bin Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *Advances in neural information processing systems*, pages 2802–2810, 2016.

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

[Mirza and Osindero, 2014] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[Patil and Mudengudi, 2011] Ujwala Patil and Uma Mudengudi. Image fusion using hierarchical pca. In *2011 International Conference on Image Information Processing*, pages 1–6. IEEE, 2011.

[Prabhakar et al., 2017] K Ram Prabhakar, V Sai Srikar, and R Venkatesh Babu. Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In *ICCV*, pages 4724–4732, 2017.

[Ronneberger et al., 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Toet and Hogervorst, 2012] Alexander Toet and Maarten A. Hogervorst. Progress in color night vision. *Optical Engineering*, 51(1):1 – 20, 2012.

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

[Zhang et al., 2017] Xiaoye Zhang, Yong Ma, Fan Fan, Ying Zhang, and Jun Huang. Infrared and visible image fusion via saliency analysis and local edge-preserving multi-scale decomposition. *Journal of the Optical Society of America A*, 34(8):1400–1410, 2017.

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