Infrared and Visible Image Fusion via Test-Time Training

Guoqing Zheng[1], Zhenqi Fu[1], Xiaopeng Lin[1], Xueye Chu[1], Yue Huang[1,2], and Xinghao Ding[1,2(✉)]

[1] School of Informatics, Xiamen University, Xiamen, China
dxh@xmu.edu.cn

[2] Institute of Artificial Intelligence, Xiamen University, Xiamen, China

Abstract. Infrared and visible image fusion (IVIF) is a widely used technique in instrument-related fields. It aims at extracting contrast information from the infrared image and texture details from the visible image and combining these two kinds of information into a single image. Most auto-encoder-based methods train the network on natural images, such as MS-COCO, and test the model on IVIF datasets. This kind of method suffers from domain shift issues and cannot generalize well in real-world scenarios. To this end, we propose a self-supervised test-time training (TTT) approach to facilitate learning a better fusion result. Specifically, a new self-supervised loss is developed to evaluate the quality of the fusion result. This loss function directs the network to improve the fusion quality by optimizing model parameters with a small number of iterations in the test time. Besides, instead of manually designing fusion strategies, we leverage a fusion adapter to automatically learn fusion rules. Experimental comparisons on two public IVIF datasets validate that the proposed method outperforms existing methods subjectively and objectively.

Keywords: Infrared image · Image fusion · Domain shift · Test-time training · Deep learning

1 Introduction

Infrared images (IR) contain thermal radiation information, and visible images (VI) have rich texture details. Infrared and visible image fusion (IVIF) aims to combine meaningful and complete information from images captured by visible and infrared sensors. As a result, the generated image contains richer information and is more favorable for subsequent computer vision tasks. IVIF techniques have been widely applied in object tracking and detection, urban security, and vehicle navigation [2,32].

In the past decades, many methods have been developed to fuse IR-VI images. Typical traditional methods are multi-scale transform-based (MST) methods and representation learning-based (RL) approaches. MST methods [25] use a specific transformation model to extract multi-scale features and manually design rules to fuse images. RL approaches include sparse representation (SR) [29], joint

Q. Liu et al. (Eds.): PRCV 2023, LNCS 14434, pp. 77–88, 2024.
https://doi.org/10.1007/978-981-99-8549-4_7

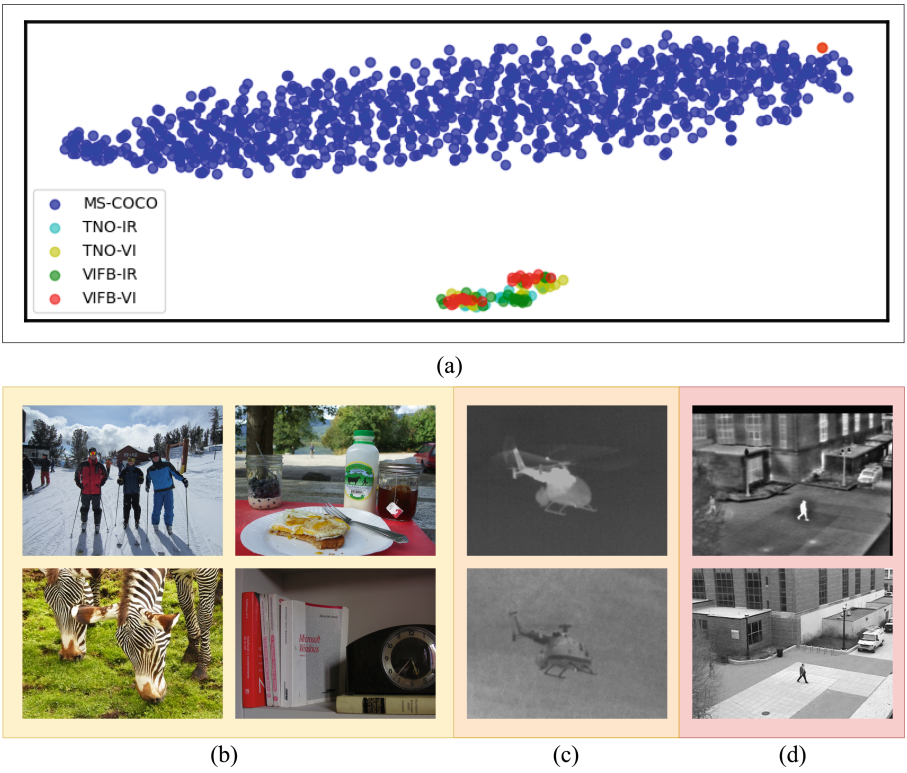

Fig. 1. (a) Visualization of image features using t-SNE [18]. Image features are generated by ResNet18 pre-trained on ImageNet [3]. These features are clustered in different centers, indicating an apparent discrepancy between the distribution of MS-COCO [11] and IVIF datasets. (b) The examples in MS-COCO. (c) and (d) are examples of IR and VI images in the TNO [24] dataset and VIFB [33] dataset, respectively.

sparse representation (JSR) [31], low-rank representation (LRR) [14], and latent low-rank representation (LatLRR) [13]. Similar to MST methods, the fusion rules in RL approaches often require manual design, which may degrade the fusion performance because source images are complex and diverse.

Recently, various deep learning-based solutions for IVIF have been presented, such as auto-encoder-based (AE) methods [6–8,26], convolutional neural network-based (CNN) methods [10], generative adversarial network-based (GAN) approaches [5], and transformer-based [23] solutions. Deep learning-based methods leverage the powerful nonlinear fitting ability of deep neural networks to make the fused images have the desired distribution. As a result, deep learning-based methods provide promising results against traditional approaches. However, deep learning-based approaches still struggle with domain shift problems, especially for AE-based solutions that train the fusion network on MS-COCO and test the model on IR-VI images. The MS-COCO and IR-VI datasets have

different characters, and the domain distribution difference between them is significant. For demonstration, we visualize the domain shift in Fig. 1.

Test-time training (TTT) techniques [22] are proposed to deal with the domain shift problem and promote the model performance for each test instance. TTT updates the network parameters before predicting. The optimization is based on specific self-supervised loss or tasks. For example, Liu et al. [15] proposed a TTT strategy that employs an auxiliary network to help the dehazing model better adapt to the domain of interest. In [4], masked auto-encoders were explored to address the one-shot learning problem, and this method improved generalization performance in various vision benchmarks.

In this paper, we propose a self-supervised TTT method that updates model parameters during the testing phase to improve the AE-based methods for IVIF. Concretely, we propose a new self-supervised loss based on a mutual attention mechanism to guide the network optimization in the test time. Moreover, we also propose a fusion adapter to automatically learn fusion rules instead of manually designing fusion strategies. Our contributions can be summarized as follows:

(1) We apply TTT strategies to improve the generalization performance of AE-based IVIF methods. A mutual attention mechanism loss is designed for self-supervised optimization.
(2) We propose a fusion adapter to learn and fuse features from different source images adaptively.
(3) Extensive experiments on two public IVIF datasets demonstrate that with a small number of iterations, the proposed method outperforms the state-of-the-art methods.

2 Method

2.1 Overall Framework

As illustrated in Fig. 2, the AE-based algorithms learn feature representations with RGB images in a self-supervised manner. Handcrafted fusion strategies are adopted in the test time to fuse IR-VI images. Note that both the encoder and the decoder are fixed during the test time. Obviously, AE-based solutions have poor generation performance because significant domain gaps exist between training and testing data. In contrast, the proposed self-supervised TTT method updates network parameters for each test sample to generate better fusion images. Besides, rather than manually design fusion strategies, we design an adapter to fuse two images. The whole network, including the encoder, the decoder, and the adapter, is optimized end-to-end during the test-time training with few iterations. Note that this paper does not focus on designing sophisticated network structures. The encoder and decoder can be any existing well-designed models.

2.2 Training and Testing

Training-Time Training. Assuming that we have a collection of large-scale dataset with training instance $X_1, X_2, ..., X_n$ drawn i.i.d from a distribution P.

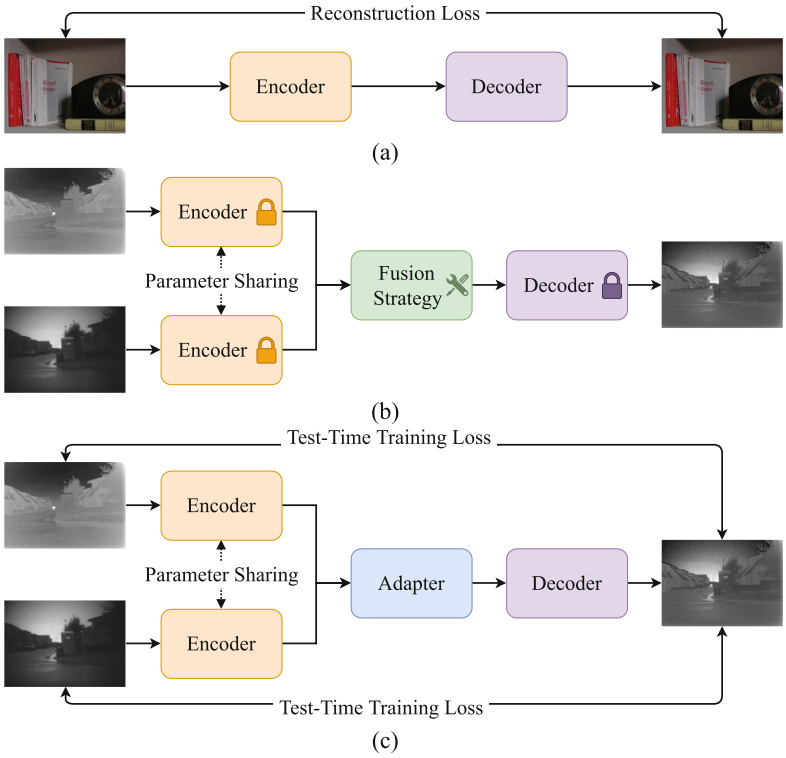

Fig. 2. Illustration of the difference between the existing AE-based IVIF methods and the proposed TTT method. (a) The AE-based solutions train the fusion network on MS-COCO and (b) test the network on IR-VI images by involving a handcrafted fusion strategy. Instead of performing (b), we propose (c) a self-supervision test-time training loss for updating the model parameters, and use a fusion adapter module to extract and fuse features of source images adaptively.

We aim to train an encoder f_θ and decoder g_θ to learn effective feature representations. Specifically, we train the encoder and decoder using a self-supervised method by minimizing the following commonly used loss function $\mathcal{L}_{tr}$:

$$g'_\theta, f'_\theta = \arg\min_{f_\theta, g_\theta} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_{tr}(X_i, g_\theta(f_\theta(X_i))), \tag{1}$$

$$\mathcal{L}_{tr} = \mathcal{L}_{pixel} + \mathcal{L}_{SSIM}, \tag{2}$$

$$\mathcal{L}_{pixel} = \left\| X_i - \hat{X}_i \right\|_F^2, \tag{3}$$

$$\mathcal{L}_{SSIM} = 1 - SSIM(X_i, \hat{X}_i), \tag{4}$$

$$\hat{X}_i = g_\theta(f_\theta(X_i))), \tag{5}$$

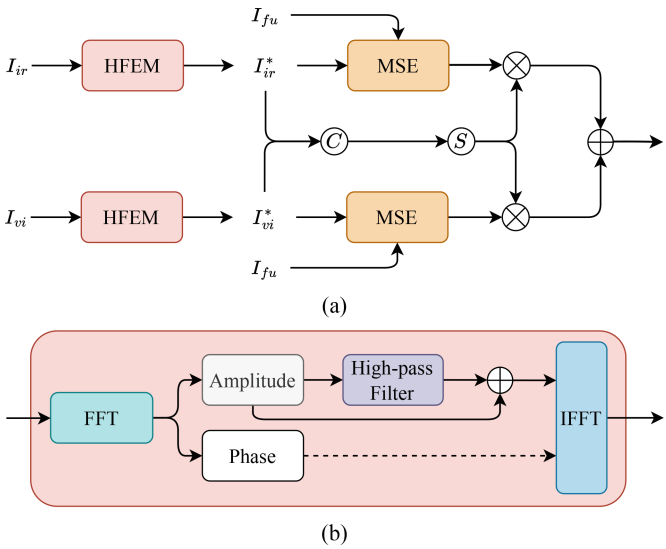

Fig. 3. (a) Overview of the proposed loss function. (b) The high-frequency enhanced module (HFEM). C refers to the concatenation operation, and S denotes the softmax function.

where $\mathcal{L}_{pixel}$ indicates the pixel level loss, $\mathcal{L}_{SSIM}$ denotes the structure similarity loss. $\|\cdot\|_F$ indicates the Frobenius norm. X_i and $\hat{X}_i$ denote the input and reconstructed images, respectively. SSIM [27] represents the structure similarity measurement.

Test-Time Training. Let I refers to the IR-VI paired dataset with test samples $(I^1_{ir}, I^1_{vi}), ..., (I^m_{ir}, I^m_{vi})$ drawn i.i.d from a distribution Q. Since the distribution Q may significantly different with P, we develop a self-supervised loss function to reduce the domain gap and promote the fusion performance. On each test input (I_{ir}, I_{vi}), we perform test-time training to minimize the following loss:

$$g^*_\theta, h^*_\theta, f^*_\theta = \arg\min_{g'_\theta, h_\theta, f'_\theta} \mathcal{L}_{TTT}(I_{ir}, I_{vi}, I_{fu}), \tag{6}$$

where $g^*_\theta, h^*_\theta, f^*_\theta$ denotes the optimized network that will be used to generate the final result. I_{fu} is the initial fusion result that obtained via the pre-trained encoder f'_θ, decoder g'_θ, and an initial adapter h_θ:

$$I_{fu} = g'_\theta(h_\theta(f'_\theta(I_{ir}), f'_\theta(I_{vi}))), \tag{7}$$

To combine meaningful information of source images, the self-supervised reconstruction loss L_{TTT} is calculated based on enhanced version of source images. As shown in Fig. 3, original IR and VI images are first inputted to a high-frequency enhanced module (HFEM) to calculate enhanced versions I^*_{ir}

and I^*_{vi}. Then, we propose a mutual attention mechanism that concatenates the I^*_{ir} and I^*_{vi} to generate two attention masks, which can be expressed as:

$$m_1, m_2 = Softmax(Concate(I^*_{ir}, I^*_{vi})), \tag{8}$$

These two attention masks force the network to pay more attention to the high-frequency information of the source images, and thus the fusion result will contain rich texture details. With the guidance of attention masks, our L_{TTT} can be defined as follows:

$$\mathcal{L}_{TTT}(I_{ir}, I_{vi}, I_{fu}) = m_1 \cdot \|I_{fu} - I^*_{ir}\|^2_F + m_2 \cdot \|I_{fu} - I^*_{vi}\|^2_F, \tag{9}$$

Note that, for each test pairs, the gradient-based optimization for Eq. (6) always starts from f'_θ, g'_θ, and h_θ. Same as [4], we discard g^*_θ, h^*_θ, f^*_θ after making a reconstructed result on each test input (I_{ir}, I_{vi}), and reset the weights to f'_θ, g'_θ, and h_θ.

Adapter. Instead of manually designing fusion rules, the proposed method develops a learnable adapter to fuse features of two source images. This is a benefit for our TTT framework, that allows the network to update parameters in the test time. The calculation of the proposed adapter h_θ is defined as follow:

$$H = Conv(ReLU(Conv(I_{co}))), \tag{10}$$

$$I_{co} = Concat(f'_\theta(I_{ir}), f'_\theta(I_{vi}))), \tag{11}$$

where H is the output feature of the adapter, I_{co} is the concatenate of features generated by the encoder, $Conv(\cdot)$ is a $2D$ convolutional layer, $ReLU(\cdot)$ is the nonlinear activation function. Our adapter has two convolutional layers and one activation function layer. During the TTT process, the adapter parameters will be updated along with the encoder and decoder to further adapt to the data distribution of the test samples.

3 Experiments

3.1 Experiment Configuration

Baseline Model. In this work, we focus on the problem of domain shifts in IVIF, especially for AE-based solutions. Therefore, we select several representation AE-based methods as our baseline models. Specifically, we choose DenseFuse [6] as our default baseline model. We also conduct experiments on other AE-based solutions [7,26] to verify the effectiveness of our proposed method in the ablation study.

Training Dataset. During the training-time training step, similar to existing AE-based networks, the encoder and the decoder are pre-trained on MS-COCO [11] dataset, which consists of more than 80,000 daily RGB images. All images are converted into grayscale versions and resized to 256×256.

Testing Datasets. This study validates the proposed method on two publicly IVIF datasets. A total of 42 pairs of images are utilized. In the TNO test dataset [24], there are 21 infrared and visible image pairs of military scenes. The remaining 21 pairs are from the VIFB [33] dataset with diverse environments, including indoor, outdoor, low-light, and over-exposure scenes.

Implementation Details. During the training-time training process, we use the same epoch, batch size, learning rate, and optimizer as in the DenseFuse [6]. For the test-time training, Adam is used as an optimizer with a learning rate of 1e-4. We update the network with five iterations for each test sample during test-time training. All the experiments are conducted on the Pytorch platform, using a GeForce RTX 3090 GPU with 24GB memory.

Evalution Metrics. For quantitative analysis, we adopt five metrics to evaluate our method and other comparative approaches, including mutual information (MI) [19], information entropy (EN) [20], sum of the correlations of differences (SCD) [1], multiscale structural similarity (MS-SSIM) [28], and visual information fidelity (VIFF) [21].

Comparison Methods. To verify the effectiveness of our proposed method, we compared the proposed with nine existing state-of-the-art algorithms, including two traditional methods: gradient transfer fusion (GTF) [16], MDLatLRR [9], and seven deep-learning-based methods: DeepFuse [8], DenseFuse [6], NestFuse [7], U2Fusion [30], DDcGAN [17], Res2Fusion [26] and TLCM [12].

3.2 Performance Comparison on TNO

The qualitative results on the TNO dataset are shown in Fig. 4. All methods can fuse the infrared image's thermal radiation information and the visible image's structure information, and our proposed method achieves the best visual quality with sharper texture details. Visually, the details of GTF, MDLatLLR, DeepFuse, U2Fusion, and TLCM are not clear enough, and the texture details of the landing gear are blurred. The fused image of DDcGAN is ambiguous and suffers from distortion. Although DenseFuse, NestFuse, and Res2Fuse can fuse more thermal radiation information, textures and details are degraded. Our method introduces TTT on the AE-based method, which preserves clear texture information and can better balance the two kinds of information of the source images. The quantitative results on the TNO dataset are shown in Table 1. As can be seen, our method achieves promising performance according to five measurements. Specifically, the comparison between our method and DenseFuse demonstrates that the proposed TTT can significantly improve the fusion performance. Besides, our method achieves the best MS-SSIM and SCD, demonstrating the superiority of the proposed method.

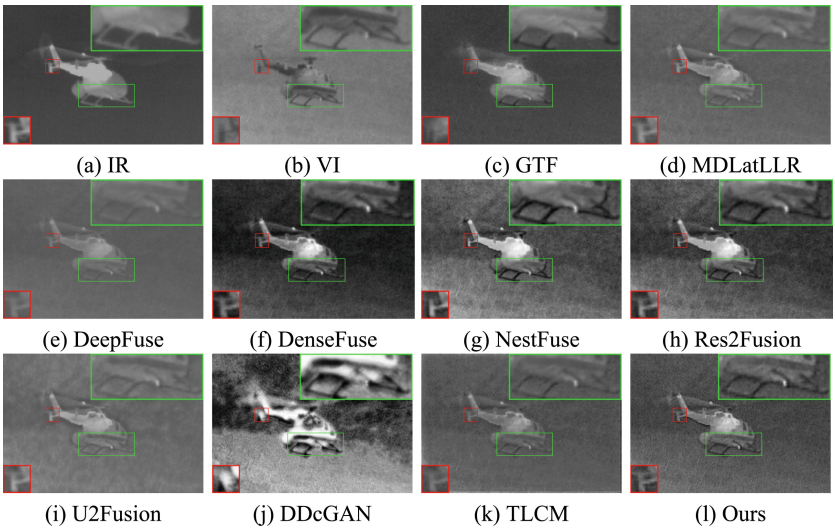

Fig. 4. Qualitative comparison of the proposed method with the state-of-the-art methods on "helicopter" on the TNO dataset. The red and green boxes show the tail and landing gear of the helicopter, respectively. (Color figure online)

Table 1. Comparisons of different methods on the TNO and VIFB dataset. Red indicates the best result, and blue represents the second best result.

Methods	TNO					VIFB				
	EN	MI	MS-SSIM	SCD	VIFF	EN	MI	MS-SSIM	SCD	VIFF
GTF	6.590	13.181	0.812	0.908	0.595	6.545	13.091	0.809	0.802	0.314
DeepFuse	6.438	12.876	0.881	1.473	0.681	6.694	13.388	0.887	1.312	0.399
DenseFuse	6.451	12.902	0.869	1.484	0.686	6.982	13.965	**0.919**	**1.497**	0.520
MDLatLLR	6.489	12.979	0.904	1.489	0.710	6.759	13.518	0.913	1.329	0.775
NestFuse	6.930	**13.961**	0.858	1.578	**0.926**	6.921	13.841	0.893	1.455	**0.861**
U2Fusion	6.468	12.936	0.899	1.459	0.681	**7.143**	**14.285**	0.907	1.461	0.832
DDcGAN	**7.479**	**14.957**	0.777	1.450	0.714	**7.484**	**14.968**	0.789	1.291	0.725
Res2Fusion	6.764	13.529	0.866	**1.727**	**0.818**	6.851	13.702	0.862	1.372	**0.852**
TLCM	**6.957**	13.914	**0.917**	1.643	0.785	6.947	13.895	0.917	1.458	0.480
Ours	6.796	13.593	**0.936**	**1.887**	0.783	7.046	14.092	**0.946**	**1.692**	0.825

3.3 Performance Comparison on VIFB

We further evaluate our method on the VIFB dataset. Qualitative results are shown in Fig. 5. As can be observed, for the brightness of the pedestrian in the red box and the detail of the car in the green box, our method has rich thermal radiation information and texture details. In addition, our method can effectively balance the global brightness of the image. The quantitative results on the VIFB dataset are shown in Table 1. The proposed method achieves the best MS-SSIM

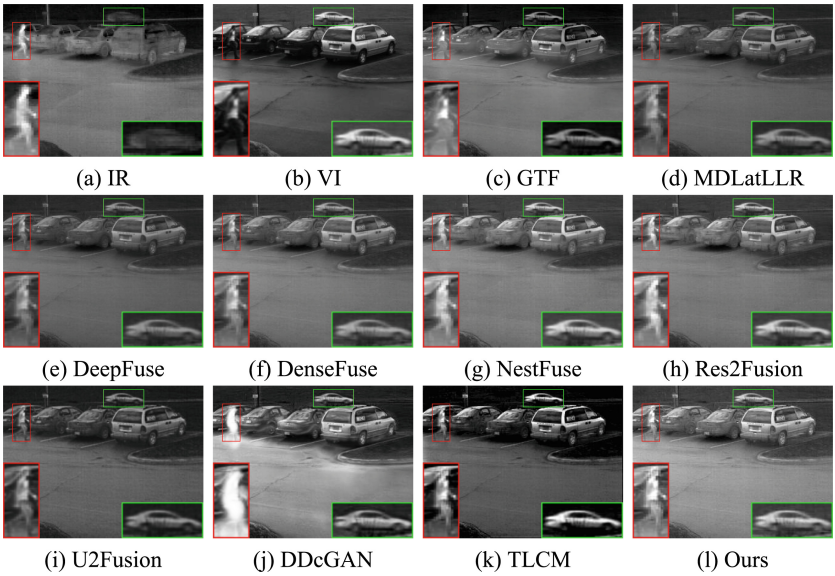

Fig. 5. Qualitative comparison of the proposed method with the state-of-the-art methods on "walking2" on the VIFB dataset. The red and green boxes show the pedestrian and car, respectively. (Color figure online)

Table 2. Ablation studies on the impact of the TTT and adapter on the TNO dataset. The best is marked in bold.

Model	TTT	Adapter	EN	MI	MS-SSIM	SCD	VIFF
DenseFuse	×	×	6.571	13.143	0.788	1.592	**0.940**
	✓	×	6.787	13.574	0.929	1.846	0.782
	✓	✓	**6.797**	**13.593**	**0.936**	**1.887**	0.783
NestFuse	×	×	**6.892**	**13.785**	0.879	1.758	**0.924**
	✓	×	6.796	13.592	**0.929**	1.849	0.787
	✓	✓	6.730	13.459	**0.929**	**1.881**	0.786
Res2Fusion	×	×	6.764	13.529	0.866	1.727	**0.878**
	✓	×	6.770	13.540	0.924	1.842	0.779
	✓	✓	**6.790**	**13.580**	**0.925**	**1.849**	0.790

and SCD scores. Our method can largely improve the performance of DenseFuse with the proposed TTT strategy.

3.4 Ablation Study

The Effectiveness of TTT. The proposed method can be added to most existing AE-based IVIF methods. Here, we choose DenseFuse, NestFuse, and

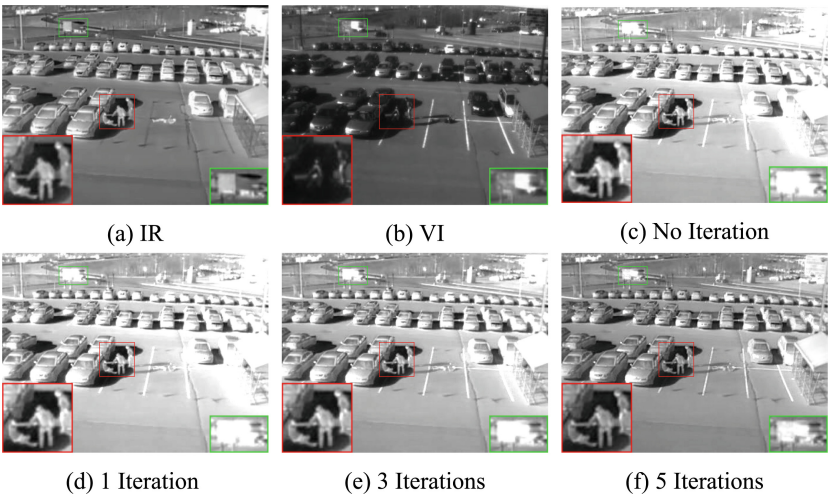

Fig. 6. Influence of the number of iterations in the TTT stage on the results of IR-VI image fusion. The quality of the fused images can be enhanced by the proposed TTT after a small number of iterations.

Res2Fusion as the baseline methods to verify the proposed method. During the training-time training phase, MS-COCO is employed to train the fusion model, and the hyperparameter settings are consistent with each original baseline method. The experiment results of different baseline methods with and without TTT are listed in Table 2. From the table, one can observe that TTT can greatly improve performance.

The Effectiveness of the Fusion Adapter. To understand the role of the adapter, we replace the adapter with feature averaging, denoted by without adapter. Test results are reported in Table 2, which indicates that the proposed adapter is better than handcrafted fusion rules. This is because our adapter is learnable and can handle diverse and complex real-world scenarios.

Ablation Studies of the Iteration. We investigate the impact of the number of iterations in the TTT stage. As shown in Fig. 6, with the iterations increase, the visual quality of the fused image gradually improves. Note that, in our experiments, continually increasing the iterations can obtain different quality outputs. But we found five updates have already achieved promising results. By comprehensively considering the run-time and performance, we set the default iteration times as five.

4 Conclusion

In this paper, we focus on developing test-time training methods to improve the generalization performance of AE-based networks. We propose a self-supervision

loss to drive the model parameters updating during the test time. This loss function is based on a mutual attention mechanism. Additionally, a fusion adapter module is proposed to adaptively fuse features of two source images. Extensive experiments and ablation studies strongly support the proposed method.

Acknowledgements. The work was supported in part by the National Natural Science Foundation of China under Grant 82172033, U19B2031, 61971369, 52105126, 82272071, 62271430, and the Fundamental Research Funds for the Central Universities 20720230104.

References

1. Aslantas, V., Bendes, E.: A new image quality metric for image fusion: the sum of the correlations of differences. Aeu-Inter. J. Electr. Commun. **69**(12), 1890–1896 (2015)
2. Das, S., Zhang, Y.: Color night vision for navigation and surveillance. Transp. Res. Rec. **1708**(1), 40–46 (2000)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. Ieee (2009)
4. Gandelsman, Y., Sun, Y., Chen, X., Efros, A.: Test-time training with masked autoencoders. Adv. Neural. Inf. Process. Syst. **35**, 29374–29385 (2022)
5. Gao, Y., Ma, S., Liu, J.: Dcdr-gan: a densely connected disentangled representation generative adversarial network for infrared and visible image fusion. IEEE Trans. Circ. Syst. Video Technol. (2022)
6. Li, H., Wu, X.J.: Densefuse: a fusion approach to infrared and visible images. IEEE Trans. Image Process. **28**(5), 2614–2623 (2018)
7. Li, H., Wu, X.J., Durrani, T.: Nestfuse: an infrared and visible image fusion architecture based on nest connection and spatial/channel attention models. IEEE Trans. Instrum. Meas. **69**(12), 9645–9656 (2020)
8. Li, H., Wu, X.J., Kittler, J.: Infrared and visible image fusion using a deep learning framework. In: 2018 24th International Conference On Pattern Recognition (ICPR), pp. 2705–2710. IEEE (2018)
9. Li, H., Wu, X.J., Kittler, J.: Mdlatlrr: a novel decomposition method for infrared and visible image fusion. IEEE Trans. Image Process. **29**, 4733–4746 (2020)
10. Li, Q., et al.: A multilevel hybrid transmission network for infrared and visible image fusion. IEEE Trans. Instrum. Meas. **71**, 1–14 (2022)
11. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
12. Lin, X., Zhou, G., Tu, X., Huang, Y., Ding, X.: Two-level consistency metric for infrared and visible image fusion. IEEE Trans. Instrum. Meas. **71**, 1–13 (2022)
13. Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., Ma, Y.: Robust recovery of subspace structures by low-rank representation. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 171–184 (2012)
14. Liu, G., Lin, Z., Yu, Y.: Robust subspace segmentation by low-rank representation. In: Proceedings of the 27th International Conference on Machine Learning (ICML 2010), pp. 663–670 (2010)

15. Liu, H., Wu, Z., Li, L., Salehkalaibar, S., Chen, J., Wang, K.: Towards multi-domain single image dehazing via test-time training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5831–5840 (2022)
16. Ma, J., Chen, C., Li, C., Huang, J.: Infrared and visible image fusion via gradient transfer and total variation minimization. Inform. Fusion **31**, 100–109 (2016)
17. Ma, J., Xu, H., Jiang, J., Mei, X., Zhang, X.P.: Ddcgan: a dual-discriminator conditional generative adversarial network for multi-resolution image fusion. IEEE Trans. Image Process. **29**, 4980–4995 (2020)
18. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. J. Mach. Learn. Res. **9**(11) (2008)
19. Piella, G.: A general framework for multiresolution image fusion: from pixels to regions. Inform. Fusion **4**(4), 259–280 (2003)
20. Roberts, J.W., Van Aardt, J.A., Ahmed, F.B.: Assessment of image fusion procedures using entropy, image quality, and multispectral classification. J. Appl. Remote Sens. **2**(1), 023522 (2008)
21. Sheikh, H.R., Bovik, A.C.: Image information and visual quality. IEEE Trans. Image Process. **15**(2), 430–444 (2006)
22. Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., Hardt, M.: Test-time training with self-supervision for generalization under distribution shifts. In: International Conference on Machine Learning, pp. 9229–9248. PMLR (2020)
23. Tang, W., He, F., Liu, Y.: Ydtr: infrared and visible image fusion via y-shape dynamic transformer. IEEE Trans. Multimedia (2022)
24. Toet, A.: The tno multiband image data collection. Data Brief **15**, 249–251 (2017)
25. Vishwakarma, A.: Image fusion using adjustable non-subsampled shearlet transform. IEEE Trans. Instrum. Meas. **68**(9), 3367–3378 (2018)
26. Wang, Z., Wu, Y., Wang, J., Xu, J., Shao, W.: Res2fusion: infrared and visible image fusion based on dense res2net and double nonlocal attention models. IEEE Trans. Instrum. Meas. **71**, 1–12 (2022)
27. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
28. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers 2003, vol. 2, pp. 1398–1402. IEEE (2003)
29. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. IEEE Trans. Pattern Anal. Mach. Intell. **31**(2), 210–227 (2008)
30. Xu, H., Ma, J., Jiang, J., Guo, X., Ling, H.: U2fusion: a unified unsupervised image fusion network. IEEE Trans. Pattern Anal. Mach. Intell. **44**(1), 502–518 (2020)
31. Zhang, Q., Fu, Y., Li, H., Zou, J.: Dictionary learning method for joint sparse representation-based image fusion. Opt. Eng. **52**(5), 057006–057006 (2013)
32. Zhang, X., Demiris, Y.: Visible and infrared image fusion using deep learning. IEEE Trans. Pattern Anal. Mach. Intell. (2023)
33. Zhang, X., Ye, P., Xiao, G.: Vifb: a visible and infrared image fusion benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 104–105 (2020)