Letter

SSL-WAEIE: Self-Supervised Learning with Weighted Auto-Encoding and Information Exchange for Infrared and Visible Image Fusion

Gucheng Zhang, Rencan Nie, and Jinde Cao, *Fellow, IEEE*

Dear editor,

Infrared and visible image fusion (IVIF) technologies are to extract complementary information from source images and generate a single fused result [1], which is widely applied in various high-level visual tasks such as segmentation and object detection [2].

Traditional fusion methods mainly include spatial domain-based methods and multi-scale decomposition-based (MSD) methods. The former ones, such as the guided filter-based methods (GF) [3] and Bayesian [4], are to produce fusion images generally by weighting local pixels or saliency of source images. MSD methods, including TE-MST [5], Hybrid-MSD [6], MDLatLRR [7], etc, first decompose the source images into multi-scale features, and further employ fusion rules to integrate these features at each level, for a reconstructed result. However, how to pixel-wisely measure the importance or fusion contribution of source images is always an open problem in these methods. Such they must elaborately design reasonable weighed strategies or fusion rules.

In recent years, deep learning has emerged as a powerful tool to perform image fusion tasks [8], [9]. Different from traditional methods, it can adaptively extract the multi-level features and automatically reconstruct the result we expected, guided by a reasonable loss. Supervised learning-based methods, such as FuseGAN [10], are mainly devoted to multi-focus image fusion because there is ground truth for each pair of training images. For our IVIF task, the most popular ways are unsupervised learning-based methods, including SeAFusion [2], FusionGAN [11], DenseFuse [12], RFN-Nest [13], DIDFuse [14], and DualFuse [15], etc, in which the network architecture is generally designed as an encoder-decoder. Similar technology is also utilized for multi-exposure image fusion [16]. However, we are not uncertain whether the features from the encoder are all the best ones. For this point, self-supervised learning-based algorithms, such as TransFuse [17] and SFA-Fuse [18], have been developed, where the encoder is designed to conduct an auxiliary task and extract the features with a prior. Nevertheless, we can also note in these methods that the features from the encoder are directly employed to reconstruct the fused result, only guided by a loss design. That is, their importance or fusion contribution not be well measured.

For contribution estimation, Nie *et al.* [19] stated a very novel idea based on information exchange. A person with less knowledge tends to learn more information from the other one with more knowledge, and vice versa, implying that this one will provide fewer contributions when they cooperatively perform a certain task. Based on this principle, the work [19] constructed a Pulse Coupled Neural network (PCNN)-based information exchange module, and applied it to perform the contribution estimation for multi-modal medical image fusion, where the fusion contribution can be easily estimated, via an exchanged information-based inverse proportional rule. Unfortunately, this module can not be optimized by the derivative-based image fusion method, due to the existing hard threshold of PCNN, and not be also trained on a large-scale dataset due to the complicated structure of a neuron.

To tackle these challenges above, in this letter, we propose a self-supervised learning-based fusion framework, named SSL-WAEIE, for the IVIF task. There are two key ideas in our method. First, we design a weighted auto-encoder (WAE) to extract the multi-level features to be fused, from source images. Second, inspired by the basic principle of information exchange in [19], we further construct a convolutional information exchange network (CIEN) to easily complete the fusion contribution estimation for source images. The main contributions of our method can be summarized as follows.

1) A novel self-supervised learning-based fusion framework. To our best knowledge, it is the first try to perform the image fusion via a convolutional neural network (CNN)-based information exchange in a manner of self-supervised learning.

2) A new network architecture. Our SSL-WAEIE consists of a WAE and a CIEN, where the WAE designed for an auxiliary task is to extract the multi-level features from source images, whereas CIEN contributes to estimating the fusion contribution mainly via a CNN-based information exchange module (IEM-CNN).

3) Hybrid losses. We propose two hybrid losses to effectively train the WAE and SSL-WAEIE, respectively.

Methodology:

Overview: Given a pair of infrared and visible images, $\mathcal{I} = (\mathbf{I}_i, \mathbf{I}_v)$, our IVIF task can be simply formulated as a weighted problem as follows

$$\mathbf{F} = \tau_{\boldsymbol{i}} \otimes \mathbf{I}_i + \tau_{\boldsymbol{v}} \otimes \mathbf{I}_v \tag{1}$$

where $\mathbf{I}_i, \mathbf{I}_v \in \mathbb{R}^{M\times N}$ denote the infrared and visible images, respectively, and $\mathbf{F} \in \mathbb{R}^{M\times N}$ is the fused result, whereas $\tau_{\boldsymbol{i}}, \tau_{\boldsymbol{v}} \in \mathbb{R}^{M\times N}$ are the fusion weights or fusion contributions to each source image, respectively, $\otimes$ is the Hadamard Product operation. Now the key to (1) is how to estimate the fusion contributions. To this end, we construct a fusion framework shown in Fig.1, which is composed of a WAE and a CIEN. Specifically, To conduct a self-supervised auxiliary task, the WAE is to extract multi-level features from $\mathcal{I}$. On the other hand, in the information exchange encoder (IEE), each IEM-CNN will perform the information exchange among pairs of features of source images at each level, to produce multi-level pairs of exchanged information. Then the contribution estimation decoder (CED) will reduct the multi-level exchanged information and finally generate the fusion contribution according to an exchanged information-based inverse proportional operation.

WAE architecture: The WAE with two Siamese branches is designed to perform an auxiliary task to the weighted reconstruction over $\mathcal{I}$. There are three convolutional blocks (CB) in each branch of the encoder, each of which sequentially includes two 2D convolutions (Conv2) layers and a ReLU function. Hence, the multi-level features of $\mathcal{I}$ can be formulated as follows.

$$\mathbf{F}_{i(l)} = \mathcal{Z}_{WAE-E}(\mathbf{F}_{i(l-1)}, \mathcal{K}_{WAE-E(l)}) \tag{2}$$

$$\mathbf{F}_{v(l)} = \mathcal{Z}_{WAE-E}(\mathbf{F}_{v(l-1)}, \mathcal{K}_{WAE-E(l)}) \tag{3}$$

where $\mathcal{Z}_{WAE-E}$ stands for the feature mapping of each CB in the encoder, whereas$\mathbf{F}_{i(l-1)}$ or $\mathbf{F}_{v(l-1)}$ is its input at the l-level, and $\mathcal{K}_{WAE-E(l)}$ is the set of filters. As outputs, $\mathbf{F}_{i(l)}$, $\mathbf{F}_{v(l)} \in \mathbb{R}^{M\times N}$ denote the features for the infrared and visible images, respectively, and $\mathbf{F}_{i(0)} = \mathbf{I}_i$, $\mathbf{F}_{v(0)} = \mathbf{I}_v$. The decoder has a similar architecture to the encoder. Differently, its function is to perform the channel reduction of inputs, completing the final image reconstruction, whereas the encoder is to extract the multi-level features from $\mathcal{I}$.

CIEN architecture:

Corresponding author: Rencan Nie.

Citation: G. C. Zhang, R. C. Nie, and J. D.Cao, "SSL-WAEIE: Self-Supervised learning with weighted auto-encoding and information exchange for infrared and visible image fusion," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 0, pp. 1–4, May 2022.

G. C. Zhang and R. C. Nie are with the School of Information Science and Engineering, Yunnan University, Kunming 650500, China (e-mail: zhang_zgc@mail.ynu.edu.cn; rcnie@ynu.edu.cn). R. C. Nie is also with School of Automation, Southeast University, Nanjing 210096, China.

J. D. Cao is with the School of Mathematics, Southeast University, Nanjing 210096, China, and Yonsei Frontier Lab, Yonsei University, Seoul 03722, South Korea. (e-mail: jdcao@seu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JAS.2022.000000

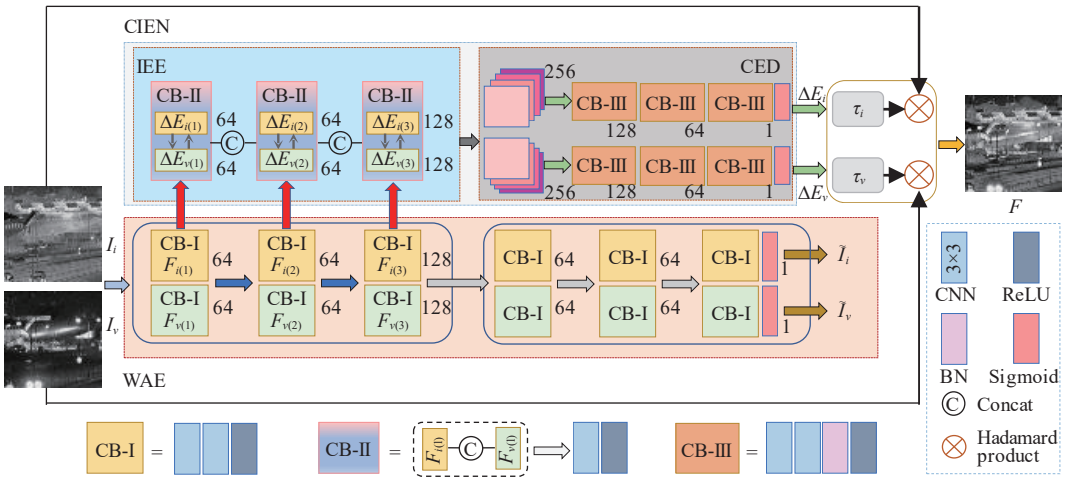

Fig. 1. Our fusion framework contains a weighted auto-encoder (WAE) and a convolutional information exchange network (CIEN), where the CIEN is composed of a information exchange encoder (IEE) and a contribution estimation decoder (CED). Note that each subnetwork is constructed by different convolutional block (CB), and there are three types of CBs, termed as CB-I to CB-III, respectively. In particular, a CB-II just corresponds to a CNN-based information exchange module (IEM-CNN).

● IEM-CNN: According to the basic principle of convolution in CNN, a feature map is from the sum of channels for the convolutions over each channel, which just is a way of information communicating among channels. Therefore, instead of PCNN in [19], we can employ 2D convolutions to simulate the mechanism of information exchange among pairs of features. In Fig. 1, we exploit three IEM-CNNs to construct the IEE.

In detail, given inputs $F_{i(l)}, F_{v(l)} \in \mathbb{R}^{C_l \times M \times N}$, we first concatenate them as a tensor of $F_{iv(l)} \in \mathbb{R}^{2C_l \times M \times N}$, then perform a 2D-convolution on $F_{iv(l)}$ by utilizing C_l pairs of convolution kernels, to produce C_l pairs of exchanged information after a ReLU function. This process can be formulated as follows

$$\Delta\mathbf{E}_{i(l)}, \Delta\mathbf{E}_{v(l)} = split(\mathcal{Z}_{IE}(\mathbf{F}_{i(l)} \circ \mathbf{F}_{v(l)}, \mathcal{K}_l)) \tag{4}$$

where $\mathcal{K}_l = (K_{i,(1,l)}, \ldots, K_{i,(C_l,l)}, K_{v,(1,l)}, \ldots, K_{v,(C_l,l)})$ is the set of filters at the l-level, and $K_{i,(x,l)}$ and $K_{v,(x,l)}$ are the x-pair of filters to produce the exchanged information between the x-pair of features $(\mathbf{F}_{i(x,l)}, \mathrm{F}_{v(x,l)})$. $\mathcal{Z}_{IE}()$ contains 2D convolutions with a kernel size of 3×3 and a ReLU function, whereas $\circ$ denotes a concatenation operation. Resorting to the *split* function, we can split a tensor into two chunks $(\Delta\mathbf{E}_{i(l)}, \Delta\mathbf{E}_{v(l)})$ with the same channels, where $\Delta\mathbf{E}_{i(l)} \in \mathbb{R}^{C_l \times M \times N}$ is the exchanged information from $\mathbf{F}_{v(l)}$ to $\mathbf{F}_{i(l)}$, vice versa for $\Delta\mathbf{E}_{v(l)}$. Compared with the IE-PCNN in [19], our IEM-CNN has several advantages: (1) It can be optimized by resorting to a derivative-based method, such that it can be trained on a large-scale dataset; (2) in a fire-new manner, it presents a very concise network architecture to complete information exchange; (3) it can perform the information exchange among pairs of multi-level features, not limited a single-scale source image.

● CED: The work [13] employs the multi-level features from the encoder to reconstruct the fused result. Similarly, we fed the multi-level exchanged information from the IEE, into our CED to produce a pair of single-scale exchanged information. As a Siamese network with the ability of channel reduction, each branch of the CED is composed of three convolution blocks, each of which includes two 2D convolutions with a kernel size of 3×3, a BatchNorm, and a ReLU function. Hence, we have

$$\Delta\mathbf{E}_i = \mathcal{Z}_{CEN}(\Delta\mathbf{E}_{i(1)} \circ \Delta\mathbf{E}_{i(2)} \circ \Delta\mathbf{E}_{i(3)}, \mathcal{K}_{CEN}) \tag{5}$$

$$\Delta\mathbf{E}_v = \mathcal{Z}_{CEN}(\Delta\mathbf{E}_{v(1)} \circ \Delta\mathbf{E}_{v(2)} \circ \Delta\mathbf{E}_{v(3)}, \mathcal{K}_{CEN}) \tag{6}$$

where $\Delta\mathbf{E}_i \in \mathbb{R}^{M \times N}$ is the single-scale exchanged information from the visible image to infrared image, similarly for $\Delta\mathbf{E}_v$. $\mathcal{Z}_{CEN}$ presents the channel reduction map of CED on a tensor, and the shared $\mathcal{K}_{CEN}$ is the set of filters in any one branch.

Now, according to the inverse proportional operation based on exchanged information, we can obtain the fusion contribution estimations as follows.

$$\tau_i = \Delta\mathbf{E}_v / (\Delta\mathbf{E}_i + \Delta\mathbf{E}_v) \tag{7}$$

$$\tau_v = \Delta\mathbf{E}_i / (\Delta\mathbf{E}_i + \Delta\mathbf{E}_v) \tag{8}$$

Training:

● Loss to WAE: The loss of the WAE is composed of two items: weighted reconstruction loss L_{W-WAE} and structural loss L_{ssim}, them are balanced by a weight λ, such that

$$L_{WAE} = L_{W-WAE} + \lambda L_{ssim} \tag{9}$$

Among them, L_{W-WAE} is to encourage the WAE to reconstruct source images in the sense of minimizing pixels.

$$L_{W-WAE} = \|\mathbf{W}_i \otimes (\widetilde{\mathbf{I}}_i - \mathbf{I}_i)\|_2^2 + \|\mathbf{W}_v \otimes (\widetilde{\mathbf{I}}_v - \mathbf{I}_v)\|_2^2 \tag{10}$$

where $\mathbf{W}_i$, $\mathbf{W}_v \in \mathbb{R}^{M \times N}$ are the weights for $\mathbf{I}_i$ and $\mathbf{I}_v$, respectively, $\mathbf{W}_i = \mathbf{I}_i / (\mathbf{I}_i + \mathbf{I}_v)$, and $\mathbf{W}_v = 1 - \mathbf{W}_i$. This weighted strategy is a simple and effective way that the encoder can extract more representation from the pixels with higher intensity, given source images. Additionally, we further employ L_{ssim} to encourage the encoder to extract more structural features from $\mathcal{I}$, such that

$$L_{ssim} = 2 - SSIM(\widetilde{\mathbf{I}}_i, \mathbf{I}_i) - SSIM(\widetilde{\mathbf{I}}_v, \mathbf{I}_v) \tag{11}$$

where $SSIM(\cdot)$ measure the structural similarity [20].

● Loss to SSL-WAEIE: The loss in our SSL-WAEIE also includes two items: the weighted fidelity loss L_{wf} and information exchange loss L_{IE}, with a regularization parameter μ.

$$L_{CIEN} = L_{wf} + \mu L_{IE} \tag{12}$$

where L_{wf} aims to produce a similarity between the fused result and each source image, which is defined as

$$L_{wf} = \|\mathbf{W}_i \otimes (\mathbf{F} - \mathbf{I}_i)\|_2^2 + \|\mathbf{W}_v \otimes (\mathbf{F} - \mathbf{I}_v)\|_2^2 \tag{13}$$

Similar to (7), $\mathbf{W}_i$ and $\mathbf{W}_v$ are beneficial to fuse more pixel intensity from $\mathcal{I}$, alleviating the luminance degeneration of the fused result. Moreover, we expect that the information exchange among pairs of features should be sufficient as far as possible because the more sufficient the information exchange between two objects is, the more accurately and easily we can measure their contributions to a certain cooperative task. To this end, for our IVIF task, we should maximize the difference among pairs of exchanged information at each level, such that

$$L_{IE} = -\frac{1}{MN} \sum\nolimits_l \|\Delta\mathbf{E}_{i(l)} - \Delta\mathbf{E}_{v(l)}\|_2^2 \tag{14}$$

Experiments:

Training details: We select 16 pairs of infrared and visible images, collected from the TNO[1] dataset, to train our SSL-WAEIE.

[1] https://figshare.com/articles/TN_Image_Fusion_Dataset/1008029

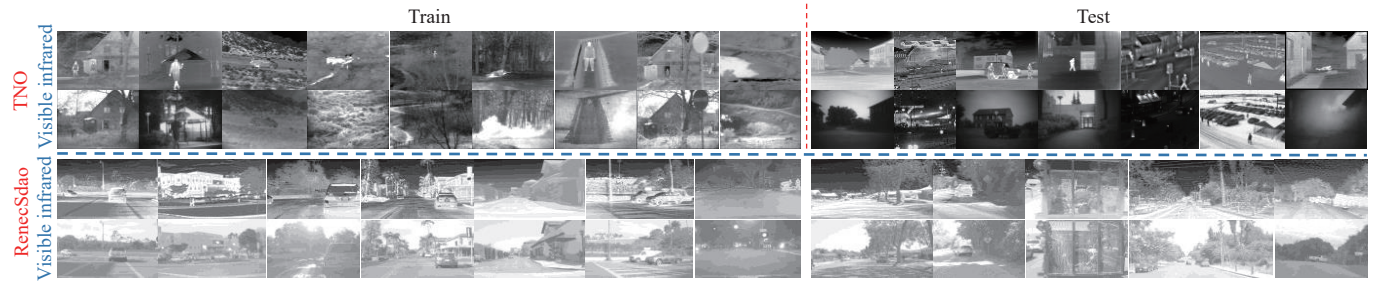

Fig. 2. Several samples of training and test datasets from two public datasets. For the TNO or RoadScene, the first row of each dataset represents the infrared images, whereas the second row denotes the visible images. Additionally, the left parts are the examples of the training samples, wheres the examples of test samples are on the right.

In Pytorch, the optimizer is set as Adam with the parameters $\beta_1 = 0.5$, $\beta_2 = 0.999$, and $\varepsilon = 1e-8$, whereas the learning rate is 2×10^{-5}. Note, however, that this learning rate for the WAE is only a tenth of what it was before, in the phase of fine-tuning. Additionally, the constant λ in the loss of WAE is empirically set to 5. On the other hand, in the testing phase, 21 pairs of source images from the TNO dataset and 20 image pairs from the RoadScene[2] dataset [21] are employed, where the parts of which are shown in Fig.2.

Setups: We compare the fusion results of our SSL-WAEIE with nine state-of-the-art methods, which contain four traditional technologies such as MDLatLRR [7], TE-MST [5], Bayesian [4], and Hybrid-MSD [6], and five deep learning-based algorithms, i.e., FusionGAN [11], DenseFuse [12], RFN-Nest [13], DIDFuse [14] and DualFuse [15]. Moreover, as four typical metrics, Normalized Mutual Information (NMI) [22], Yang's Metrics (Q_Y) [23], Gradient-based Metric (Q_{ABF}) [24], and Information Fidelity Criterion (IFC) [25] are employed to quantitatively evaluate the results.

Results: As shown in Fig. 3, the results from FusionGAN, DenseFuse, DualFuse, and TE-MST, suffer from obscure edges for the infrared objects. On the other hand, although RFN-Nest and Bayesian preserve acceptable background from the visible image, their results show low contrast due to luminance degradation. Additionally, MDLatLRR produces obvious artifacts, especially in Fig. 4. DIDFuse and Hybrid-MSD achieve good results, whereas the details and luminance of the targets in the red rectangle are still defective. Compared to these methods, our SSL-WAEIE not only depicts a significant improvement in luminance but also retains more texture details furthest.

Table 1 illustrates that our method gives the best quantitative results in terms of NMI, Q_Y, and IFC, on the TNO dataset, compared with nine competitors, whereas it ranks second for Q_{ABF}. The same conclusion, on RoadScene dataset, also can be drawn from Table 2.

Parameter analysis: The regularization parameter μ in our loss plays an important role to train our network. Hence, we vary it from 9 to 11 to investigate its influence on the fusion performance related to each metric. Table 3 shows that as μ increases, all metrics increase before 10. They then decrease on the whole and fall into an oscillation period. Hence, we take the value of μ as 10 for our SSL-WAEIE.

Ablation study:

● Ablation to network: Discarding the WAE and IEE, respectively, our SSL-WAEIE degenerates into two versions: No WAE and No IEE. In the first one, the decoder of the WAE will be destroyed, such that it will be not pre-trained via our auxiliary task. In the other one, the multi-level features from the decoder of the WAE will be directly concatenated and fed into CEN to generate the fused result. Table 4 shows that No IEE presents the worst results in terms of each metric, whereas SSL-WAEIE is the best one. Therefore, WAE and IEE both provide significant contributions to our method, however, IEE is more important. Moreover, let IEE perform the information exchange on only the last level features from WAE, such that our IEE turns to a single-level one (SL-IEE) and only has an IEM-CNN. We can see that SL-IEE, in terms of each metric, is superior to No IEE, whereas it is inferior to the IEE. Hence, instead of single-level information exchange, the multi-level one in our IEE is more beneficial to produce a good result, this also implies that our WAE can extract good multi-level features rather than trivial solutions.

● Ablation to loss. In Table 4, $L_1 = L_f$ is the baseline loss, through discarding the weighted strategy to$L_2 = L_{wf}$ in our L_{CIEN}. We can see that L_2 is superior to L_1 in terms of each metric, whereas the quantitative results are further improved by our L_{CIEN}. Therefore, the weighted strategy and the constraint to the information exchange in

[2] https://github.com/jiayi-ma/RoadScene

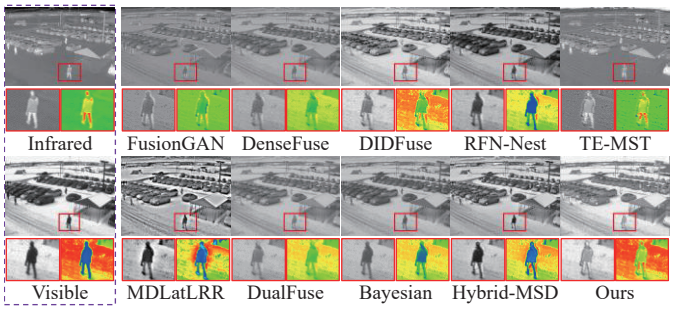

Fig. 3. An example of visual comparison among different methods on TNO dataset, where the results of our method have a significant advantage over other algorithms.

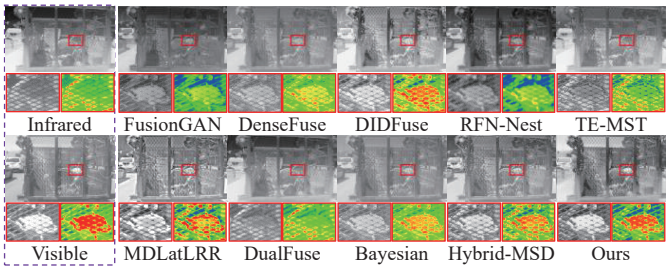

Fig. 4. An example of visual comparison among different methods on RoadScene dataset. Compared to other algorithms, our proposed method gives the best visual result.

Table 1. Average Quantitative Results on TNO for Different Methods

Metrics	NMI ↑	Q_Y ↑	Q_{ABF} ↑	IFC ↑
FusionGAN [11] (2019)	0.3852	0.6075	0.2993	2.2958
DenseFuse [12] (2018)	0.3794	0.6742	0.3542	2.6165
DIDFuse [14] (2021)	0.3730	0.6067	0.3785	2.2139
RFN-Nest [13] (2021)	0.3509	0.6363	0.3356	2.3313
TE-MST [5] (2020)	**0.5770**	**0.8070**	0.5044	2.9937
MDLatLRR [7] (2020)	0.2523	0.7423	0.4817	**3.0380**
DualFuse [15] (2021)	0.3663	0.7021	0.4076	2.6076
Bayesian [4] (2020)	0.4110	0.7147	0.3641	2.7767
Hybrid-MSD [6] (2016)	0.4207	0.8037	**0.5427**	2.8068
Ours	**0.5826**	**0.8320**	**0.5412**	**3.6330**

Best and second results are shown in green and blue **bold** fonts, respectively.

Table 2. Average Quantitative Results on RoadScene for Each Method

Metrics	NMI ↑	Q_Y ↑	Q_{ABF} ↑	IFC ↑
FusionGAN [11] (2019)	0.4365	0.5308	0.3056	1.3666
DenseFuse [12] (2018)	0.4402	0.6389	0.3617	1.7094
DIDFuse [14] (2021)	0.4430	0.6733	0.4345	1.8302
RFN-Nest [13] (2021)	0.4369	0.6121	0.3485	1.6660
TE-MST [5] (2020)	**0.5708**	0.7409	0.4761	2.646
MDLatLRR [7] (2020)	0.3484	0.7503	0.4968	**2.6157**
DualFuse [15] (2021)	0.3888	0.4800	0.2936	1.3317
Bayesian [4] (2020)	0.4689	0.7079	0.4118	1.9616
Hybrid-MSD [6] (2016)	0.4668	**0.8218**	**0.5364**	2.5806
Ours	**0.6617**	**0.8653**	**0.5001**	**2.9144**

Best and second results are shown in green and blue **bold** fonts, respectively.

Table 3. Quantitative Results in Terms of Each Metric for Different Regularization Parameters in L_{CIEN}, on TNO.

Parameter (μ)	NMI ↑	Q_Y ↑	Q_{ABF} ↑	IFC ↑
$\mu = 9$	0.5402	0.8070	0.5177	3.3803
$\mu = 9.5$	0.5476	0.8161	0.5220	3.4300
$\mu = 10$ (Ours)	**0.5826**	**0.8320**	**0.5412**	**3.6330**
$\mu = 10.5$	0.5422	0.8285	0.5339	3.5110
$\mu = 11$	0.5467	0.8223	0.5289	3.4377

Best results are shown in green **bold** fonts.

Table 4. Quantitative Results for the Ablation Study on TNO.

Ablation	Methods	NMI ↑	Q_Y ↑	Q_{ABF} ↑	IFC ↑
Network	No WAE	0.4580	0.8126	0.5167	2.7935
	No IEE	0.5081	0.7734	0.4412	2.5825
	SL-IEE	0.5325	0.8297	0.5313	3.2468
	IEE (Ours)	**0.5826**	**0.8320**	**0.5412**	**3.6330**
Loss	$L_1 = L_f$	0.3782	0.7514	0.4683	2.5998
	$L_2 = L_{wf}$	0.5433	0.8122	0.5201	3.1900
	L_{CIEN} (Ours)	**0.5826**	**0.8320**	**0.5412**	**3.6330**

Best results are shown in green **bold** fonts.

L_{CIEN} are both beneficial to improve the fusion performance. It is worthy to note that the implementation of information exchange in our method is not only via the loss L_{IE}, but more resulted from the network architecture with IEM-CNN. That is to say that our fusion framework also can be performed if there is no L_{IE}, whereas this regularization can further improve the fusion performance resulting from the mechanism of information exchange.

Conclusions: In this letter, we propose a novel self-supervised learning-based fusion network for infrared and visible images. A WAE is designed to perform an auxiliary task for the weighted reconstruction over source images. Such that we can further employ the multi-level features from the encoder of WAE to perform the exchanged information-based fusion contribution estimation in CIEN. Particularly, according to the principle of information exchange, we employ CNN to specifically construct an information exchange module, such that our CIEN can easily complete the fusion contribution estimation for source images. Moreover, we employ the weighted strategy and the constraint to the information exchange to design a hybrid loss to effectively train our SSL-WAEIE. Extensive experiments on two public datasets verify the superiority of our method to other state-of-art competitors.

Acknowledgments: This work was supported by National Natural Science Foundation of China (61966037, 61833005, and 61463052), China Postdoctoral Science Foundation (2017M621586), and Postgraduate Science Foundation of Yunnan University (2021Y263).

References

[1] G. Bhatnagar and Q. J. Wu, "A fractal dimension based framework for night vision fusion," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 1, pp. 220–227, 2018.

[2] L. Tang, J. Yuan, and J. Ma, "Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network," *Information Fusion*, 2022.

[3] S. Li, X. Kang, and J. Hu, "Image fusion with guided filtering," *IEEE Trans. Image processing*, vol. 22, no. 7, pp. 2864–2875, 2013.

[4] Z. Zhao, S. Xu, C. Zhang, J. Liu, and J. Zhang, "Bayesian fusion for infrared and visible images," *Signal Processing*, vol. 177, Article No. 107734, 2020.

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

[6] Z. Zhou, W. Bo, L. Sun, and M. Dong, "Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with Gaussian and bilateral filters," *Information Fusion*, vol. 30, 2016.

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

[8] Y. Liu, X. Chen, Z. Wang, Z. J. Wang, R. K. Ward, and X. Wang, "Deep learning for pixel-level image fusion: Recent advances and future prospects," *Information Fusion*, vol. 42, pp. 158–173, 2018.

[9] Y. Liu, L. Wang, J. Cheng, C. Li, and X. Chen, "Multi-focus image fusion: A survey of the state of the art," *Information Fusion*, vol. 64, pp. 71–91, 2020.

[10] X. Guo, R. Nie, J. Cao, D. Zhou, L. Mei, and K. He, "FuseGAN: Learning to fuse multi-focus image via conditional generative adversarial network," *IEEE Trans. Multimedia*, vol. 21, no. 8, pp. 1982–1996, 2019.

[11] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Information Fusion*, vol. 48, pp. 11–26, 2019.

[12] H. Li and X.-J. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Processing*, vol. 28, no. 5, pp. 2614–2623, 2018.

[13] H. Li, X.-J. Wu, and J. Kittler, "RFN-Nest: An end-to-end residual fusion network for infrared and visible images," *Information Fusion*, vol. 73, pp. 72–86, 2021.

[14] Z. Zhao, S. Xu, C. Zhang, J. Liu, P. Li, and J. Zhang, "DIDFuse: Deep image decomposition for infrared and visible image fusion," *arXiv preprint arXiv: 2003.09210*, 2020.

[15] Y. Fu and X.-J. Wu, "A dual-branch network for infrared and visible image fusion," in *2020 25th Int. Conf. on Pattern Recognition (ICPR)*, 2021: IEEE, pp. 10675–10680.

[16] K. Ram Prabhakar, V. Sai Srikar, and R. Venkatesh Babu, "Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs," in *Proc. the IEEE international conference on computer vision*, 2017, pp. 4714–4722.

[17] L. Qu *et al.*, "TransFuse: A Unified Transformer-based Image Fusion Framework using Self-supervised Learning," *arXiv preprint arXiv: 2201.07451*, 2022.

[18] F. Zhao, W. Zhao, L. Yao, and Y. Liu, "Self-supervised feature adaption for infrared and visible image fusion," *Information Fusion*, vol. 76, pp. 189–203, 2021.

[19] R. Nie, J. Cao, D. Zhou, and W. Qian, "Multi-source information exchange encoding with pcnn for medical image fusion," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 31, no. 3, pp. 986–1000, 2020.

[20] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[21] H. Xu, J. Ma, Z. Le, J. Jiang, and X. Guo, "Fusiondn: A unified densely connected network for image fusion," in *Proc. the AAAI Conf. on Artificial Intelligence*, 2020, vol. 34, no. 07, pp. 12484–12491.

[22] R. Nie, C. Ma, J. Cao, H. Ding, and D. Zhou, "A total variation with joint norms for infrared and visible image fusion," *IEEE Trans. Multimedia*, 2021.

[23] C. Yang, J.-Q. Zhang, X.-R. Wang, and X. Liu, "A novel similarity based quality metric for image fusion," *Information Fusion*, vol. 2, no. 9, pp. 156–160, 2008.

[24] C. a. Xydeas and V. Petrovic, "Objective image fusion performance measure," *Electronics letters*, vol. 36, no. 4, pp. 308–309, 2000.

[25] H. R. Sheikh, A. C. Bovik, and G. De Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Trans. image processing*, vol. 14, no. 12, pp. 2117–2128, 2005.