Integration of Bi-dimensional Empirical Mode Decomposition With Two Streams Deep Learning Network for Infrared and Visible Image Fusion

Manoj Kumar Panda[1], Badri N Subudhi[1], T.Veerakumar[2] and Vinit Jakhetiya[3]
[1]Department of Electrical Engineering, Indian Institute of Technology Jammu, Jammu and Kashmir, India
[2]Department of Electronics and Communication Engineering, National Institute of Technology Goa, Ponda, Goa, India
[3]Department of Computer Science & Engineering, Indian Institute of Technology Jammu, Jammu and Kashmir, India
manojpanda2014@gmail.com, subudhi.badri@gmail.com, tveerakumar@yahoo.co.in, vinit.jakhetiya@iitjammu.ac.in

***Abstract*—Image fusion is a technique that combines the complementary details from the images captured from different sensors into a single image with high perception capability. In the fusion process, the significant details from different source images are combined in a meaningful way. In this article, we propose a unique and first effort of infrared and visible image fusion technique with bi-dimensional empirical mode decomposition (BEMD) induced VGG-16 deep neural network. The proposed BEMD strategy is incorporated with a pre-trained VGG-16 network that can effectively handle the vagueness of infrared and visible images and retain deep multi-layer features at different scales on the frequency domain. A novel fusion strategy is proposed here to analyze the spatial inter-dependency between these features and precisely preserve the correlative information from the source images. The minimum selection strategy is explored in the proposed algorithm to keep the standard details with reduced artifacts in the fused image. The competency of the proposed algorithm is estimated using qualitative and quantitative assessments. The efficiency of the proposed technique is corroborated against fifteen existing state-of-the-art fusion techniques and found to be efficient.**

***Index Terms*—Multi-scale decomposition, Deep neural network, Infrared image, Visible image**

I. Introduction

Image fusion is a process that combines the correlative information from the source images and produces a fused image with lesser artifacts. The resultant image is apparent to be more extensive and suitable for human or machine perception. The source images are generated from infrared (IR) and visual sensors or a single camera with unlike imaging measures. It is observed that the use of both IR and visible images is effective for the vision-based system due to the complementary and pervasive characteristics [1]. The use of fusion technologies is successfully applied in the field of image enhancement, object detection, object recognition, surveillance, etc. [1]. The crucial requirement of an effective image fusion scheme is to extract the meaningful features from the source images and fuse the same to produce a fused image with lesser noise. In this regard, different fusion techniques are developed by the researchers.

In the last two decades, the popular geometrical transformation based image fusion techniques have been popularly used and developed: Laplacian pyramid [2], contrast pyramid [2], discrete wavelet transform [2], non-subsampled contourlet transform [3] and curvelet transform [3] etc. These techniques extract the features from the different source images at multi-scale & multi-level and combine the same by using an appropriate fusion rule. The fused image is generated based on inverse transformation. However, these fusion schemes are susceptible to mis-registration and are also incapable of preserving sufficient details from the source images. Adjacent to these fusion schemes, sparsity and dictionary learning-based fusion schemes [3]–[5] are also applied successfully in state-of-the-art techniques. These techniques can represent the source images over a complete dictionary by utilizing sparse coefficients. These sparse coefficients are amalgamate accurately using the relevant fusion rules to produce the fused image. However, dictionary learning-based approaches are very much time-consuming, and the resultant images obtained by the same have more artifacts. In recent years, several deep learning-based image fusion algorithms [6]–[9] are getting it's popularity due to effective multi-scale features representation ability. In these techniques, the deep features are extracted from the source images at different levels and utilized to construct the fused images. But these techniques are incapable of making full use of deep features and face difficulties retaining the source details efficiently.

In this paper, we have developed an effective IR, and visible fusion algorithm based on bi-dimensional empirical mode decomposition (BEMD) strategy persuaded VGG-16 deep neural network. The BEMD mechanism decomposes the source images into several intrinsic mode functions ($IMFs$) at different frequency bands. The proposed BEMD strategy with VGG-16 architecture explores features in-depth on frequency domain at various levels and can handle the high uncertainty in the source images. The proposed deep multi-level fusion strat-

Authors would like to acknowledge MeitY, India for providing the financial support with grant no. 4(16)/2019-ITEA dated 28/02/2020.

egy constructs the weight maps to preserve the correlative data accurately from the images of different sensors and provides a detailed fusion map. The minimum selection strategy among these detailed maps retains the standard information and reduces the superfluous data. The proposed model provides a fused image with lesser artifacts for a pair of input IR and visible images.

The rest of the paper is organized as follows. The proposed fusion model with graphical illustration is discussed in section II. Section III comprises the experimental result analysis of the proposed and existing fusion models. The conclusion of the work is summarized in section IV.

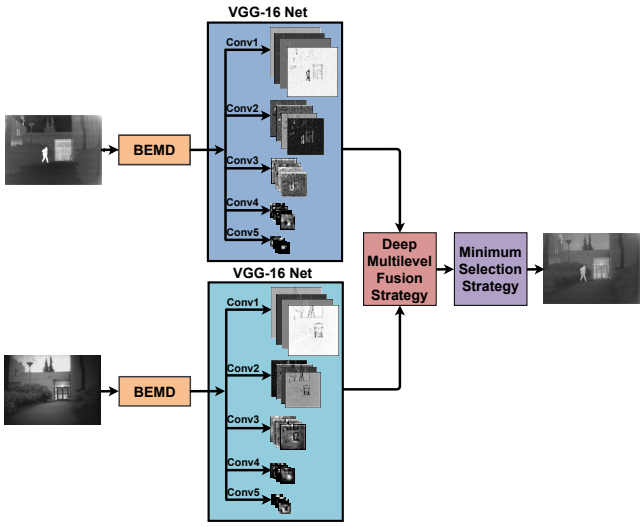

Fig. 1: Block diagram of the proposed algorithm.

II. The Proposed Fusion Algorithm

In this paper, we have proposed an effective IR and visible image fusion technique which can be further used for several real-life applications. The source images have high uncertainty and may possess camera noise. Therefore, it is a quite challenging task to extract the meaningful features from them and propagate them into a fused image with reduced artifacts. In this regard, the proposed Bi-dimensional empirical mode decomposition (BEMD) strategy is integrated with the VGG-16 deep learning framework that can extract multi-scale deep features from the source images. The proposed novel fusion strategy retains significant visual details at a multi-level to produce a fused image with lesser artifacts. In this article, the source images are represented as I_j, $j \in \{1, 2\}$, 1 corresponds to the IR, and 2 corresponds to the visible image, and the fused image is expressed as F. The graphical exposition of the proposed scheme is narrated in Fig. 1. Here, the BEMD block is combined with the VGG-16 network to preserve the deep multi-level visual details at various scales. In the proposed deep multi-level fusion process, the obtained features are combined and then the minimum selection strategy is used to produce a resultant image with complementary details and lesser noise.

A. Bi-dimensional empirical mode decomposition

The empirical mode decomposition (EMD) [10] technique is popularly used in the signal and image processing domain to decomposes any signal into finite oscillatory components. It is an adaptive algorithm and relevant for stationary as well as non-stationary signal analysis. The extracted oscillatory components from the signal are named as intrinsic mode function (IMF). It is observed that the EMD mechanism plays an important role in one-dimensional signal analysis. The EMD mechanism is further extended and utilized for two-dimensional signal or image analysis and is known as bi-dimensional empirical mode decomposition (BEMD) [11]. The BEMD strategy extracts the $IMFs$ from the source images by utilizing the sifting process [11] and can be described as;

$$\begin{aligned} IMF_j^n(x,y) = \{&IMF_j^1(x,y), IMF_j^2(x,y), \\ &\cdots\cdots, IMF_j^{N-1}(x,y), R_j^N(x,y)\}; \\ \forall n = 1, 2, \cdots\cdots, N-1, N. \end{aligned} \tag{1}$$

Here $IMF_j^n(x,y)$ indicates the $IMFs$ for the images of thermal and visual sensors. $R_j^N(x,y)$ denotes the residue bands of source images.

B. The fusion of the intrinsic mode functions

We have introduced the BEMD strategy to decompose the source images into N numbers of intrinsic mode functions at multi-scale with different frequency bands. To retain the maximum details in the fused image with reduced noise, it is necessary to extract the deep features at different levels from the $IMFs$ and combine them accurately. Therefore, we have proposed a unique deep multi-level fusion strategy with VGG-16 (deep learning architecture) [12] that provides the multi-scale and multi-level visual characteristics of the considered scene. The considered VGG-16 architecture consists of convolutional layers, max-pooling layers, and a rectified linear unit (ReLU) as an activation function with five convolutional blocks. The convolutional layers are used to retain the spatial information of the source images, and max-pooling layers are used as a down-sampling operation. The presence of the ReLU function in the network makes it faster and efficient.

In the proposed algorithm, we have used three $IMFs$ except residual bands of source images to extract the deep features as shown in Fig. 1. Let us assume that the $\psi_j^{t,1:c}$ deep features are extracted by the VGG-16 network with t^{th} convolutional block consisting of c channels, $t \in \{1, 2, 3, 4, 5\}$. $\psi_j^{t,1:c}(x,y)$ indicates the contents at position (x,y). Initially, the corresponding $IMFs$ from the source images are given to the deep VGG-16 network separately to retain the deep features. Subsequently, at each block, we have utilized the sum of the absolute difference (SAD) operator to map from features space to image space named as initial action map ϕ_j^t and can be given as;

$$\phi_j^t(x,y) = SAD(\psi_j^{t,1:c}(x,y)). \tag{2}$$

To make the proposed scheme prosperous to mis-registration, we have considered a center sliding window $u \times u$ in the ϕ_j^t.

A block-based average (BBA) operator in the center sliding window is used to obtain the final action map $\hat{\phi}_j^t$ and can be calculated as;

$$\hat{\phi}_j^t(x,y) = \frac{\sum_{a=-\frac{u-1}{2}}^{\frac{u-1}{2}} \sum_{b=-\frac{u-1}{2}}^{\frac{u-1}{2}} \phi_j^t(x+a, y+b)}{u^2}. \quad (3)$$

For the larger value of u, the proposed algorithm is more prosperous to mis-registration. However, at the same time, small-scale details may be lost, which are essential for multi-modal image fusion. Hence, in this proposed work, we have kept the size of the center sliding window to 3×3.

To precisely preserve the complementary information from the source images, we have developed an initial weight map w_j^t by using the normalization operator in the $\hat{\phi}_j^t$. The w_j^t in the range [0,1] can be determined as;

$$w_j^t(x,y) = \frac{\hat{\phi}_j^t(x,y)}{\sum_{m=1}^{j} \hat{\phi}_m^t(x,y)}. \quad (4)$$

As we know, the max-pooling layer with stride $s = 2$ in the VGG-16 architecture reduces the size of the input feature to $1/s$ times. Hence, the bi-cubic interpolation is utilized in the w_j^t to generate the final weight map $\hat{w}_j^t$ where the size of $\hat{w}_j^t$ same as the source image size.

The initial detail maps idm^t generated from the source images and final weight maps to retain the high strength details and remove low strength details. Now, we have five pairs of final weight maps, and for each pair of final weight maps, the initial detail map can be calculated as;

$$idm^t(x,y) = \sum_{z=1}^{j} \hat{w}_z^t(x,y) \times I_z(x,y). \quad (5)$$

The final detail map fdm is obtained by using maximum selection strategy among these initial detail maps to preserve sharp details and can be calculated as;

$$fdm(x,y) = max[idm^t(x,y)], \quad (6)$$

likewise, we acquired various final detail maps from the corresponding $IMFs$.

C. Fused image generation

To generate the fused image F, we have utilized a minimum selection strategy among these final detail maps to preserve the standard information and reduce the redundant data. The fused image is obtained as;

$$F(x,y) = min[fdm^n(x,y)], \quad (7)$$

where n indicates the number of $IMFs$ pairs.

III. Experimental Results Analysis

The efficacy of the proposed algorithm is tested on all considered IR and visible image pairs of the "TNO" benchmark database [13] with several challenging scenes: illumination variation, smoke, occluded object, and non-uniform lighting conditions, etc. The efficiency of the proposed scheme is validated both qualitatively and quantitatively. For page constraint, the visual demonstrated result is shown on few source image pairs, and quantitative analysis is shown on all the image pairs of the "TNO" benchmark database. The proposed scheme is run on a $Core$ i7 system with 16 GB RAM and python programming using Keras framework along Tensorflow backend. The achievement of the proposed technique is evaluated by comparing the results obtained by it against fifteen existing state-of-the-art techniques: cross bilateral filter (CBF) [14], weighted least square (WLS) [15], convolutional sparse representation (CSR) [4], ratio of low-pass pyramid (RP) [3], RP with sparse representation (RP-SR) [3], latent low-rank representation (LatLRR) [16], morphological component analysis based on convolutional sparsity (CS-MCA) [17], Fuzzy edge [18], Joint SR with saliency detection (JSRSD) [5], saliency detection (SD) [19], convolutional neural network (CNN) [20], deep neural network (DNN) [21], Fusion based on generative adversarial network (FGAN) [8], image fusion based on CNN (IFCNN) [9] and residual fusion network (RFN) [22]. This section is further split into qualitative assessment, quantitative assessment, and ablation study.

A. Qualitative Assessment

The original images acquired from the visual and the thermal sensors along with the results obtained by the proposed and different considered state-of-the-art techniques: CBF, RP, RP-SR, Fuzzy edge, RFN, and DNN are presented in Figs. 2. It may be observed that the resultant images procured by the different techniques used for comparison: CBF, RP, RP-SR, and Fuzzy edge have produced many artifacts and cannot retain significant details in the fused image, as shown in the red rectangle highlighted region on different image. The outcomes of the RFN technique have blurred details with more noise. Due to ringing artifacts around the edge details, the significant features are not clearly visible or highlighting non-required details in the fused image by the DNN technique. However, the results obtained by the proposed technique have maximum details with lesser artifacts.

B. Quantitative Assessment

The evaluation of fusion performance is difficult due to irrelevant variation in visual demonstrated fusion results obtained by different fusion schemes. Therefore, in this paper we have used the four most suggested fusion metrics: mutual information for the discrete cosine features (FMI_{dct}) [23], amount of artifacts added during the fusion process (N_{abf}) [23], average structure similarity index $(SSIM_a)$ [23], and average edge preservation index (EPI_a) [23]. The performance of any fusion algorithm is better if the FMI_{dct}, $SSIM_a$, and EPI_a values are higher with lower N_{abf} value.

Table I encapsulates the average quantitative measures of the proposed and state-of-the-art fusion schemes where the best values are indicated in bold. From this Table, it may be observed that the proposed scheme attained higher accuracy against fifteen existing fusion algorithms. The fused images acquired by the proposed algorithm strongly correlate with

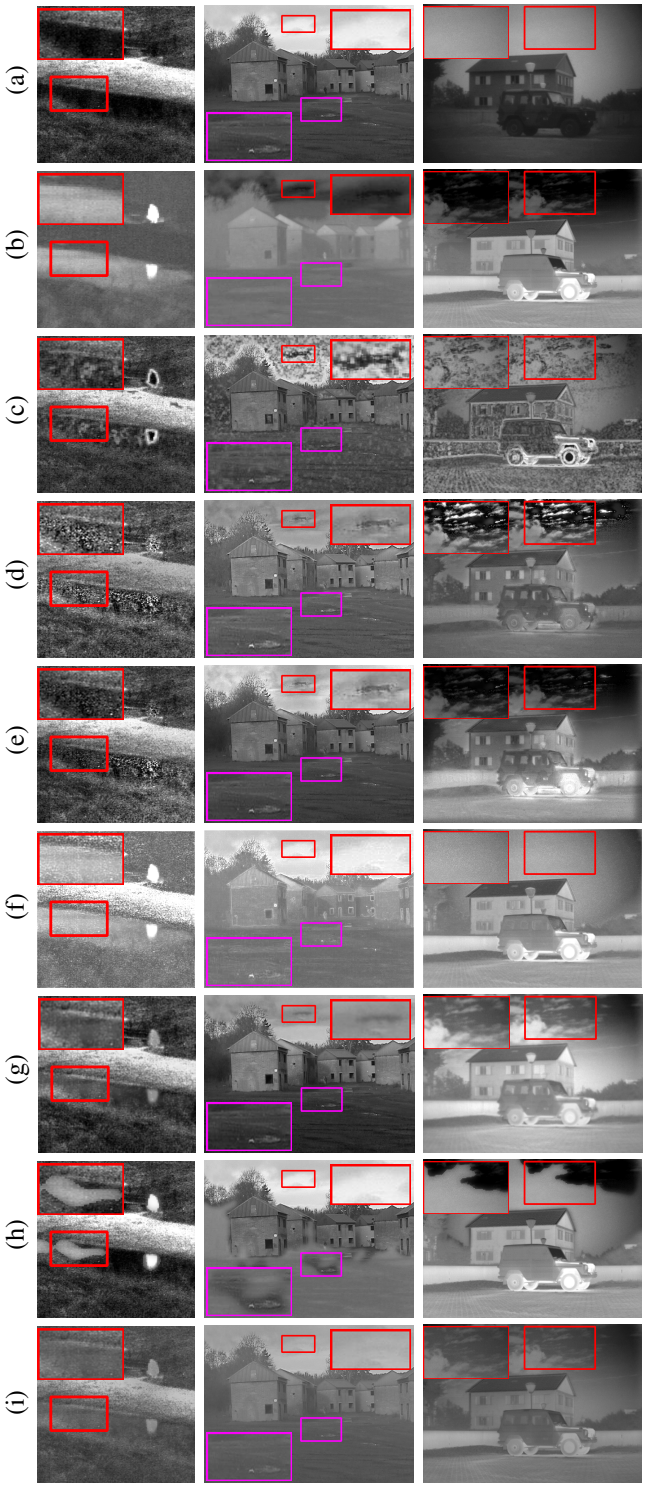

Fig. 2: Visual analysis of results on the Bench, Octec, and Marne image (from left to right). From top to bottom: (a) Visible images, (b) IR images, results of (c) CBF, (d) RP, (e) RP-SR, (f) Fuzzy edge, (g) RFN, (h) DNN and (i) Proposed schemes, respectively.

TABLE I: Quantitative comparison with existing image fusion algorithms for all the pairs of source images

Quantitative measurements /Algorithms	Avg.FMI_{dct}	Avg.N_{abf}	Avg.$SSIM_a$	Avg.EPI_a
CBF [14]	0.26309	0.31727	0.59957	0.57240
WLS [15]	0.33102	0.21257	0.72360	0.67837
CSR [4]	0.34640	0.01958	0.75335	0.71130
RP [3]	0.28210	0.22677	0.68424	0.64488
RP-SR [3]	0.27930	0.21444	0.67385	0.63737
LatLRR [16]	0.33817	0.01596	0.76486	0.76223
CS-MCA [17]	0.35841	0.06680	0.72964	0.69154
Fuzzy edge [18]	0.31052	0.28250	0.60635	0.66744
JSRSD [5]	0.14253	0.34657	0.54127	0.47473
SD [19]	0.27030	0.13430	0.72897	0.66774
CNN [20]	0.35269	0.13280	0.71372	0.68444
DNN [21]	0.36658	0.02324	0.70852	0.68552
FGAN [8]	0.36335	0.06706	0.65384	0.68470
IFCNN [9]	0.37378	0.17959	0.73186	0.73767
RFN [22]	0.29669	0.07288	0.69949	0.68864
Proposed	**0.39962**	**0.00149**	**0.77671**	**0.77909**

source images and contain fewer artifacts as compared to the existing state-of-the-art techniques.

C. Ablation Study

We conduct ablation studies to see the performance of the proposed scheme against alteration of the different components in the proposed technique. For this study, we replaced the proposed scheme for BEMD with: Proposed algorithm with two $IMFs$, Proposed algorithm with three $IMFs$ and residual bands and Proposed algorithm with three $IMFs$ and without residual bands. The performance are shown In Table II. It may be observed from this table that the proposed scheme retain maximum information in the fused image. Similar experiment is carried out for ablation study with: Two-stream VGG-16-Deep multi-level fusion strategy (Without any decomposition strategy), Laplacian decomposition [3]-Two-stream VGG-16-Deep multi-level fusion strategy, Tikhonov optimization [24]-Two-stream VGG-16-Deep multi-level fusion strategy, and BEMD-Two-stream VGG-16-Deep multi-level fusion strategy. The results are shown in Table III. It may be observed from this table that the proposed scheme retain maximum information. We also have studied the effects of entropy (EN) [18] and mutual information (MI) [18] with different window size. The studies for EN is provided in Fig. 3 and study for MI is provided in Fig. 4. From this Figs., it may be concluded that with the increasing values of window size u, the EN and MI measures are reduced, which degrade the quality of the fused image and retain fewer details from the source images. Hence, in this proposed work, we have kept the size of the center sliding window to 3×3 to get a higher values of EN and MI.

TABLE II: Ablation study of proposed algorithm with and without residual bands

Experiments	FMI_{dct}	N_{abf}	$SSIM_a$	EPI_a
Proposed algorithm with two $IMFs$	0.28366	**0.00301**	0.87356	0.95109
Proposed algorithm with three $IMFs$ and residual bands	0.28373	0.00325	0.87310	0.95150
Proposed algorithm with three $IMFs$ and without residual bands	**0.28401**	0.00316	**0.87371**	**0.95249**

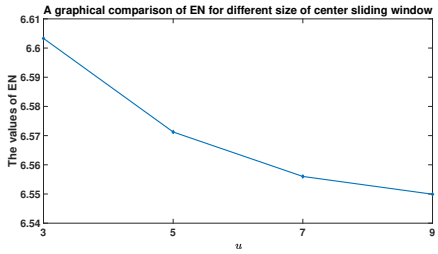

Fig. 3: Graphical illustration of entropy for different size of center sliding window.

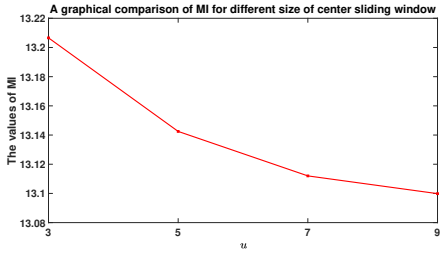

Fig. 4: Graphical illustration of mutual information for different size of center sliding window.

TABLE III: Ablation study of proposed algorithm with different image decomposition strategy.

Evaluation measures /Decomposition strategies	FMI_{dct}	N_{abf}	$SSIM_a$	EPI_a
Two-stream VGG-16-Deep multi-level fusion strategy (Without any decomposition strategy)	0.38507	0.00477	**0.68092**	0.71582
Laplacian decomposition-Two-stream VGG-16-Deep multi-level fusion strategy	0.39259	0.00533	0.68016	0.71836
Tikhonov optimization-Two-stream VGG-16-Deep multi-level fusion strategy	0.39060	0.00512	0.68050	0.71845
BEMD-Two-stream VGG-16-Deep multi-level fusion strategy	**0.39544**	**0.00445**	0.68084	**0.72181**

IV. Conclusion

In this paper, we have proposed an efficient image fusion scheme to perpetuate the correlative data from the source images to the fused image with lesser noise. The proposed BEMD strategy is integrated with a VGG-16 deep neural architecture that can learn a mapping from image space to feature space at multi-scale with different levels. The proposed multi-level fusion strategy; investigates the spatial inter-dependency among these features and accurately acquires the complementary information from the source images. The minimum selection strategy in the proposed scheme produce the fused image with reduced artifacts. The efficacy of the proposed technique is corroborated using qualitative and quantitative assessment against fifteen existing fusion schemes. It is observed that the fused images attained by the proposed algorithm have a strong correlation with the source images and higher accuracy than the state-of-the-art fusion techniques.

References

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

[2] P. Jagalingam and A. V. Hegde, "Pixel level image fusion—a review on various techniques," in *Proceedings of the 3rd World Conference on Applied Sciences, Engineering and Technology*, 2014, pp. 1–8.

[3] Y. Liu, S. Liu, and Z. Wang, "A general framework for image fusion based on multi-scale transform and sparse representation," *Information Fusion*, vol. 24, pp. 147–164, 2015.

[4] Y. Liu, X. Chen, R. K. Ward, and Z. J. Wang, "Image fusion with convolutional sparse representation," *IEEE Signal Processing Letters*, vol. 23, no. 12, pp. 1882–1886, 2016.

[5] C. Liu, Y. Qi, and W. Ding, "Infrared and visible image fusion method based on saliency detection in sparse domain," *Infrared Physics & Technology*, vol. 83, pp. 94–102, 2017.

[6] H. Li, X. J. Wu, and J. Kittler, "Infrared and visible image fusion using a deep learning framework," in *Proceedings of the 24th International Conference on Pattern Recognition*, 2018, pp. 2705–2710.

[7] H. Li and X. Wu, "Densefuse: A fusion approach to infrared and visible images," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2614–2623, 2019.

[8] 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.

[9] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, and L. Zhang, "IFCNN: A general image fusion framework based on convolutional neural network," *Information Fusion*, vol. 54, pp. 99–118, 2020.

[10] J. C. Nunes, Y. Bouaoune, E. Delechelle, O. Niang, and P. Bunel, "Image analysis by bidimensional empirical mode decomposition," *Image and Vision Computing*, vol. 21, no. 12, pp. 1019–1026, 2003.

[11] T. Veerakumar, B. N. Subudhi, and S. Esakkirajan, "Empirical mode decomposition and adaptive bilateral filter approach for impulse noise removal," *Expert Systems with Applications*, vol. 121, pp. 18–27, 2019.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2015.

[13] A. Toet *et al.*, "TNO image fusion dataset," *Figshare. data2014. https: // figshare.com/articles/TN_Image_ Fusion Dataset/1008029.*, 2014 (accessed on Aug.-16, 2021).

[14] B. S. Kumar, "Image fusion based on pixel significance using cross bilateral filter," *Signal, Image and Video Processing*, vol. 9, no. 5, pp. 1193–1204, 2015.

[15] J. Ma, Z. Zhou, B. Wang, and H. Zong, "Infrared and visible image fusion based on visual saliency map and weighted least square optimization," *Infrared Physics & Technology*, vol. 82, pp. 8–17, 2017.

[16] H. Li and X. J. Wu, "Infrared and visible image fusion using latent low-rank representation," *arXiv preprint arXiv:1804.08992*, 2019.

[17] Y. Liu, X. Chen, R. K. Ward, and Z. J. Wang, "Medical image fusion via convolutional sparsity based morphological component analysis," *IEEE Signal Processing Letters*, vol. 26, no. 3, pp. 485–489, 2019.

[18] M. K. Panda, B. N. Subudhi, T. Veerakumar, and M. S. Gaur, "Edge preserving image fusion using intensity variation approach," in *Proceedings of the IEEE Region 10 Conference*, 2020, pp. 251–256.

[19] D. P. Bavirisetti and R. Dhuli, "Two-scale image fusion of visible and infrared images using saliency detection," *Infrared Physics & Technology*, vol. 76, pp. 52–64, 2016.

[20] Y. Liu, X. Chen, J. Cheng, H. Peng, and Z. Wang, "Infrared and visible image fusion with convolutional neural networks," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 16, no. 03, pp. 1–20, 2018.

[21] Y. Liu, X. Chen, H. Peng, and Z. Wang, "Multi-focus image fusion with a deep convolutional neural network," *Information Fusion*, vol. 36, pp. 191–207, 2017.

[22] 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.

[23] M. K. Panda, B. N. Subudhi, T. Veerakumar, and M. S. Gaur, "Pixel-level visual and thermal images fusion using maximum and minimum value selection strategy," in *Proceedings of the IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security*, 2020, pp. 1–6.

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