LEARNING ATTENTION GUIDED DEEP MULTI-SCALE FEATURE ENSEMBLE FOR INFRARED AND VISIBLE IMAGE FUSION

Paper ID: 92

ABSTRACT

Infrared and visible image fusion refers to integrating series of images acquired from different sensors, obtaining a more comprehensive image. It plays a vital role in a variety of computer vision tasks. Most existing approaches utilize simple methods to transform source images into feature domain, and fuse the intermediate features with addition or concatenation strategy. However, these roughly feature extraction methods and incomplete fusion rules can not meet the need of challenging image fusion tasks, existing inadequate feature extraction and degeneration of details. In this paper, we propose a learning attention guided deep multi-scale feature ensemble with two well designed modules targeting to aforementioned issues. The coarse-to-fine scheme is introduced in feature extractor module to refine the deep features with different scales. Besides, a learning edge-guided attention mechanism is proposed to avoid the degeneration of details information and get rid of undesirable artifacts. Finally, we conducted the experiments on two datasets. Extensive results demonstrate the superiority of the proposed method, outperforming state-of-art methods quantitatively and qualitatively.

***Index Terms*—** Infrared and visible image, image fusion, deep learning, dilated convolution

1. INTRODUCTION

Multi-modality image fusion plays a critical role in computer vision community, due to it can provide high informative and reliable images for intriguing high-level vision tasks, e.g., object detection [1], pedestrian re-identification and semantic segmentation [2]. Among them, the combination of visible and infrared image has incomparable advantages. Commonly, visible images contain abundant details information with high resolution. However, under the darkness or severe environment, visible images often present bad visual performance, owing to the insufficient lighting or villainous weather. Complementarily, discriminative thermal radiation based infrared image has been invented. It is free from the outside brightness, and receives the thermal rays emitted from different objects for imaging. But the limitation is that thermal image only characterizes the general profile with low resolution, texture details of object can not be presented appropriately resulting from the properties of infrared imaging. Therefore, it is worthwhile to integrate the visible and infrared image into a single image, increasing human visual understanding and further serve for military and civil fields, e.g., supervisory control and autonomous driving.

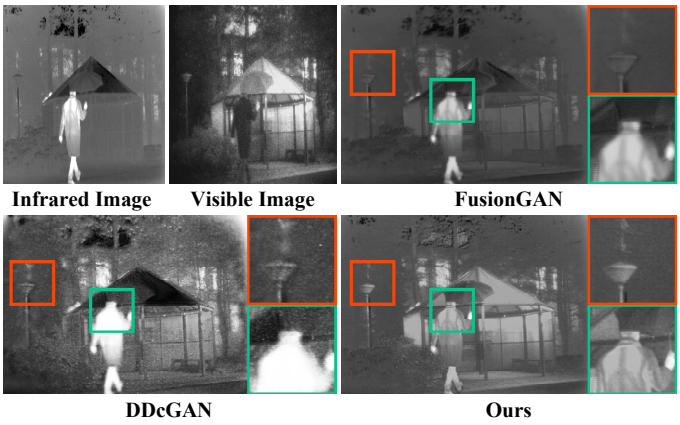

Fig. 1. Schematic illustration of the proposed method by make comparison with the two state-of-art deep learning based methods FusionGAN [3] and DDcGAN [4]. Clearly, our method not only well preserve the detail information, but also reduce the artifacts.

In the following, we will give a brief review on infrared and visible image fusion methods and summarize our contributions at the end of this section.

1.1. Related Works

The taxing point of image fusion is how to extract the characteristics from two source images and merge them to reconstruct an informative result. To this end, a great deal of image methods have been proposed, targeting to designs various effective feature extraction strategies and appropriate fusion rules. All these methods can be simply divided into two categories, *i.e.,* traditional methods and deep learning methods.

Traditional methods have their roles to play in image fusion field. Liang *et al.* [5] proposed a method for multiple fuse by formulating a tensor and used SVD to decompose the images. Kim *et al.* [6] introduced a novel dictionary learning method based on patch clustering and PCA, this method achieved the good fusion performance and removed redundancy of the learned dictionary. Nencini *et al.* [7] employed curvelet transform to separate source images into different

wavelet domain while extracting the abundant information more efficiently. Kim *et al.* [6] introduced a novel dictionary learning method based on patch clustering and principal component analysis. Ma *et al.* [8] initially adopt Gaussian and rolling guidance filers to decompose the source images into four scales in terms of base layers and detail layers, then handled the base layers with advanced visual saliency map, finally added detail layers into base layers utilizing the weight least square optimization. These methods achieved promising performance. However, these traditional approaches attempt manually designed rules, making the methods become more complex and time-consuming.

Recently, convolutional neural networks (CNNs) have been widely used in infrared and visible image fusion. Ma *et al.* [3] utilized generative adversarial network (GAN) to fuse the two source images, and preserved the main intensity from the infrared image and detail information in visible image. Xu *et al.* [4] adopted the encoder-decoder network initially, and enhanced the roughly fused image with dual conditional discriminators. Li *et al.* [9] trained a deep learning network based on the CNN layers and dense block, achieving good performance in assessment. However, these CNN based methods can not deal with detail degeneration from the feature map efficiently, moreover, the lack of elaborately designed fusion rules may also bring undesirable artifacts and halos.

1.2. Contributions

Based on a large number of experiments, we solve the task of infrared and visible image fusion by considering two taxing points, *i.e.,* details degradation and limited feature extraction. A Learning attention guided deep multi-scale feature ensemble with two well-designed modules is proposed. Specifically, it extracts multi-scale features with a coarse-to-fine densely extraction and fuses them guided by a edge based attention mechanism to avoid the degradation of details. Using the feature compensation reconstruction network, we can achieve an informative and visual-friendly fusion results. Our contributions can be summarized as follows.

- We develop a coarse-to-fine multi-scale feature extractor which employs three densely connected dilation paths to perform feature extraction. Based on this scheme, the combination of dilated convolution and dense connection embrace multi-scale reception fields, making great use of image features.
- We exploit an edge-guided attention based feature fusion mechanism, which ensures the detail information consistency between source images and the fused results, preserving the details and reducing noises or undesirable artifacts simultaneously.
- Extensive results on two datasets demonstrate the superiority of our method, outperforming state-of-the-art methods quantitatively and qualitatively.

2. THE PROPOSED FUSION METHOD

In this section, we give a detailed introduction of the proposed method from three main aspects: multi-scale feature extractor, learning attention based feature fusion as well as feature compensation reconstruction. To begin with, the input infrared and visible images are separately fed to feature extractor to acquire a series of intermediate features. Then applied a learning based module to calculate the attention maps, combing with the intermediate features to fused together. Finally, reconstruct the fused features by deconvolution with three skip connections from the feature extractor, compensating the detail lose further.

2.1. Deep Multi-scale Feature Extractor

For infrared and visible image fusion task, one import issue is extracting abundant features to represent the input images. The methods of feature extraction usually push forward a immense influence on fused result. Previous approaches simply designed feature extractor without consider contextualized information, which may introduce artifacts in the fused result. Therefore, we proposed a contextualized dilated feature extractor to obtain deep feature at multiple scales with different receptive fields. Besides, integrating dense block in each scale to guarantee the extracted features are more sufficient.

To begin with, the network transforms the infrared and visible image into feature space via the convolution operation which can be denoted as f_{ir_i} $(i = 1, ..., 16)$ and f_{vis_i} $(i = 1, ..., 16)$, respectively. Then these features are fed into three dilated convolution layers. These dilated convolution weighted according to the step size of the dilated factor, leading to increase the receptive field without changing the resolution of the image. Each dilated path consists of three convolutions with the same kernel size 3×3. Different paths use their typical receptive fields to provide more precise complementary information and receptive fields of the first layer in each path are 3×3, 5×5 and 7×7. Additionally, to preserve deep features as much as possible in feature extraction module, we added dense connection in each dilated path. Each layer's output is cascaded as the input of the next layer. The total feature from the feature extraction module is denoted as $f^e_{ir\_i}$ $(i = 1, ..., 64)$ and $f^e_{vis\_i}$ $(i = 1, ..., 64)$, respectively.

The densely dilated feature extraction module not only captures the information from different receptive fields but also makes sure all the salient features can be preserved.

2.2. Learning Guided Attention Feature Fusion

Attention mechanisms have been successfully applied in many computer vision tasks since it can capture the region of interest in the visual scene. The main target of infrared and visible image fusion is to preserve more detail information and get rid of the undesirable artifact. Therefore, we designed cross domain edge maps to trained the attention mechanism,

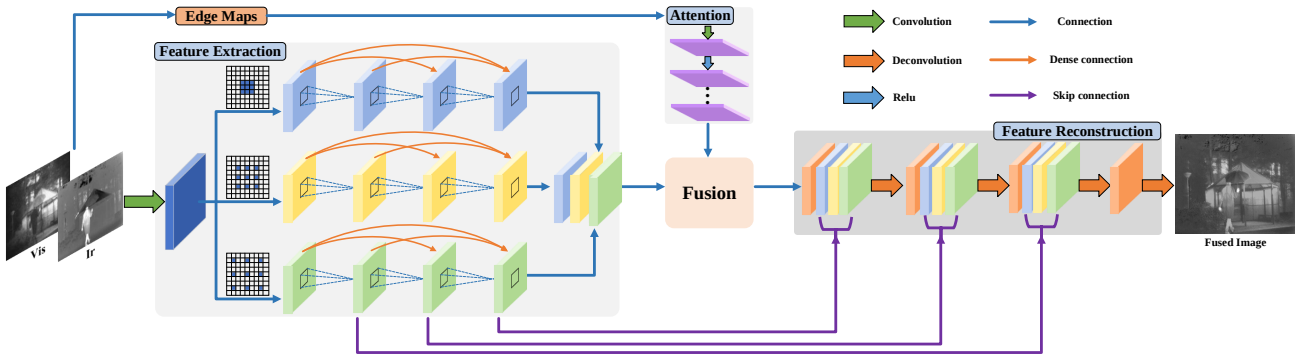

Fig. 2. Architecture of the proposed infrared and visible image fusion framework.

and obtain the edge-guided deep salient features to fuse them in feature domain. Specifically, we obtained the image edge

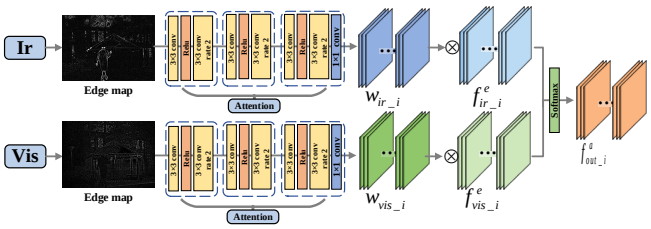

Fig. 3. The detailed procedure of edge-guide attention based fusion mechanism. *rate* 2 denotes dilated convolution operator with a dilation rate of two.

maps by two steps. Supposing input image of gray-scale with the size $m \times n$ denoted by $\mathbf{u}$, the gradient $\mathcal{G}$ is defined as:

$$\mathcal{G}(\mathbf{u}) = \sum_{i=1}^{mn} \sqrt{(\nabla_i^h \mathbf{u})^2 + \nabla_i^v \mathbf{u})^2}, \tag{1}$$

where $\nabla_i^h \mathbf{u} = \mathbf{u}_i - \mathbf{u}_{a(i)}$ and $\nabla_i^v \mathbf{u} = \mathbf{u}_i - \mathbf{u}_{b(i)}$, respectively. ∇^h, ∇^v express as the liner operator by calculating the horizontal and vertical first-order differences. $u_{a(i)}, u_{b(i)}$ represent the nearest neighbor pixel which locate at right and below the source pixel i.

Then to further make gradient information more remarkable, we proposed a edge operator $\mathcal{S}$ which defined as:

$$\begin{cases} \mathcal{G}(u) = \max\limits_{i=1,\ldots,m-1} \left(\mathcal{G}(u)(i,:), \mathcal{G}(u)(i+1,:)\right), \\ \mathcal{S}(\mathcal{G}(u)) = \max\limits_{j=1,\ldots,n-1} \left(\mathcal{G}(u)(:,j), \mathcal{G}(u)(:,j+1)\right), \end{cases} \tag{2}$$

where i and j represent the pixel index of horizontal and vertical direction of the gradient image, respectively.

Subsequently, feeding the infrared and visible enhanced edge maps into the attention mechanism to generate feature weights denoted as $W_{ir\_i}(i = 1, ..., 64), W_{vis\_i}(i = 1, ..., 64)$. Therefore, the edge-guided attention fused feature $f_{out\_i}^a (i = 1, ..., 64)$ is calculated by:

$$f_{out\_i}^a = softmax(\sum_{i=1}^{k}(f_{ir\_i}^e W_{ir\_i} + f_{vis\_i}^e W_{vis\_i})). \tag{3}$$

The intermediate feature maps are multiplied by the attention weights and summed up together to generate the fused feature maps. Specific architecture is shown in Fig. 3. And the final fused image will be reconstructed to image domain in the following module which the input total feature maps are $f_{out\_i}^a$(i=1,...,64).

2.3. Feature Compensation Reconstruction

Image reconstruction aims at transforming the intermediate feature maps from the feature space to image space by deconvolution layers. However, this roughly deconvolution operation may lose vital information to some extents. For this purpose, we added three skip connections from the feature extractor module to the corresponding deconvolution layers, making compensate for the information lose further.

Let $f_{ir\_i}^{conv\_k}(k = 2, 3, 4)$, $f_{ir\_i}^{conv\_k}(k = 2, 3, 4)$ denote the compensate infrared and visible feature obtained from the feature extractor module and sum the different dilated feature map together. Then these compensate features merge with the corresponding deconvolutional layer by element-wise summation throw the skip connection.

$$f_{out_i}^{dconv\_k-1} = f_{ir\_i}^{conv\_k} + f_{vis\_i}^{conv\_k}(k = 2, 3, 4). \tag{4}$$

Here, $f_{out_i}^{dconv\_k}(1, 2, 3)$ expresses as the total deconvolutional feature maps in the corresponding layer. Finally, we get the fused result from the feature reconstruction module.

3. EXPERIMENTS

3.1. Training Details

We trained the proposed network on MSCOCO dataset, choosing 25000 images as the training images and 1500 images as validation data. We cropped them into size 256×256. Learning rate is set as 1×10^{-4} and the batch size is 16.

Fig. 4. Visual comparison among different methods on Kaptein-1123 images which comes from TNO dataset. As shown, our method performs better in preserving detail texture, especially in the zoomed-in patches.

For multi-modality image fusion, there is no ground truth for network to execute supervised/unsupervised learning. Besides, sufficient registered images are merely obtained in real world scenery. Therefore, we firstly train our feature extractor and feature reconstruction module to learn the interconversion of image domain and feature domain. Inspired by GAN, we suppose the aforementioned network as a generator and add a discriminator at the end of network to guide the generator produce more natural images. The generator and discriminator are alternate iteration. To constrain the reconstructed structure error and pixel error, we set the training loss as the combination of $\mathcal{L}_{\text{pixel}}$, $\mathcal{L}_{\text{SSIM}}$ and $\mathcal{L}_{\text{GAN}}$ with weights γ and λ. Moreover, our total loss function expressed as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pixel}} + \gamma\mathcal{L}_{\text{SSIM}} + \lambda\mathcal{L}_{\text{GAN}}. \tag{5}$$

Where $\mathcal{L}_{\text{SSIM}}$ represents the structural similarity operation, $\mathcal{L}_{\text{pixel}}$ calculate the Euclidean distance between the input and the output image. $\mathcal{L}_{\text{GAN}}$ measures the mean square error between the input and output. Fig. 5 and Fig. 6 demonstrate the effectiveness of training scheme between GAN and original network. We can see that the training loss with GAN decreases faster than the original network in Fig. 5 and its fused result is closer to natural image in Fig. 6. Therefore, it is reasonable to introduce generative and adversarial scheme in our training phase.

Subsequently, we train the attention mechanism with the pre-trained feature extractor and reconstruction module. Through the well trained attention mechanism, we are able to get the edge-guided weight maps to learn the salient details of source images.

3.2. Results Analysis

We make comparison with five state-of-the-art methods, including transform based method CBF [10], total variation method GTF [11], sparse coding method JSRSD [12] and two deep learning methods, FusionGAN [3] and DDcGAN [4].

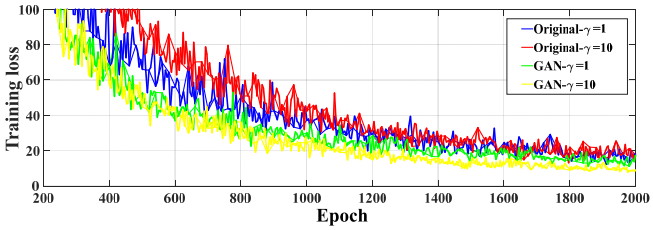

Fig. 5. Training loss between the proposed GAN based scheme and the original network. It is clearly that the proposed GAN based training scheme decreases faster than the other one and achieves a lower loss value.

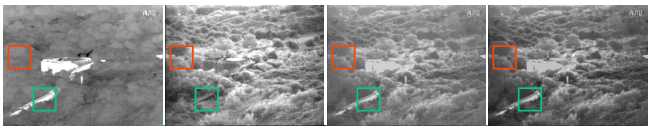

Fig. 6. From left to right: visible image, infrared image, our results without and with discriminator in training phase. Clearly, the discriminator guided fused result is closer to a natural image, especially in red and green boxes.

Experiments are conducted on two benchmarks, *i.e.,* TNO which contains 3 video sequence and 25 image pairs.

Qualitative Comparisons Fig. 4 and Fig. 7 illustrate two pairs of results on two datasets respectively. Compared with other methods, the proposed method has three significant advantages. Firstly, our method preserved abundant details. As shown in Fig. 4, the contours of human body in our result is more clearly than the others since we adopt three different groups of dilated convolution to extract multi-scale features. Secondly, our results is able to avoid undesirable artifacts. In Fig. 4, little noise merges around the tree which is benefit from attention mechanism. Lastly, our results contains more salient details. In Fig. 7, visual result not only shows the external structure of the car but also reconstructs its internal

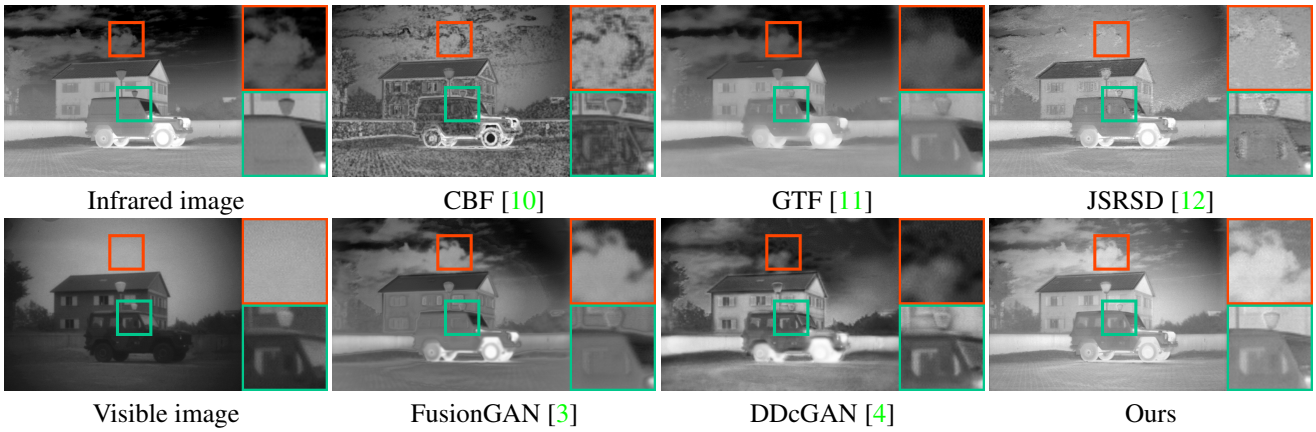

Fig. 7. Visual results comparison between different methods on Marne_04 from TNO dataset. It can be seen that only our method preserve the cloud clearly without noises degrade and artifacts, especially in red frame.

information. What's more, our method restores the shape of cloud completely.

Table 1. The quantitative results on three sequence datasets from TNO. And the best results are highlighted in bold.

	Metric	CBF	GTF	JSD	FG	DD	Ours
Nato	MI	1.43	1.99	1.48	2.42	2.34	**2.44**
	CC	0.57	0.46	0.58	0.62	0.67	**0.67**
	VIF	0.30	0.36	0.17	0.22	**0.43**	0.40
	$Q^{AB/F}$	0.34	0.37	0.20	0.18	0.22	**0.39**
	SCD	1.34	0.96	1.56	1.12	0.89	**1.77**
	SSIM	0.41	0.43	0.25	0.53	0.35	**0.55**
Tree	MI	1.37	1.67	1.61	1.62	2.19	**2.59**
	CC	0.64	0.61	0.59	0.62	0.66	**0.70**
	VIF	0.26	0.41	0.20	0.25	0.52	**0.53**
	$Q^{AB/F}$	0.21	0.35	0.16	0.18	0.27	**0.31**
	SCD	1.20	0.71	1.02	1.65	1.08	**1.67**
	SSIM	0.30	0.44	0.23	0.23	0.37	**0.50**
Duine	MI	1.26	1.23	0.80	1.77	1.67	**1.93**
	CC	0.67	0.52	0.55	0.53	0.28	**0.72**
	VIF	0.37	0.46	0.16	0.38	0.41	**0.51**
	$Q^{AB/F}$	0.34	**0.42**	0.11	0.24	0.37	0.39
	SCD	1.25	0.73	1.21	0.96	1.07	**1.93**
	SSIM	0.52	0.52	0.18	0.35	0.36	**0.53**

Quantitative Comparisons We employ six evaluation metrics, including mutual information (MI) [13], correlation coefficient (CC) [14], visual information fidelity (VIF) [15], $Q^{AB/F}$ [16], the sum of the correlations of differences (SCD) [17] and structure similarity (SSIM) [18] to further verify the effectiveness of the proposed method on two benchmarks. As shown in Fig. 8 and Tab. 1, our method generates the largest average value on the all six metrics in different datasets. What's more, its performance is more stable than others. The large MI and $Q^{AB/F}$ metrics indicate that our method transferred more considerable details and edge information from source images, while the great VIF metric represents our results are more friendly to human visual system. Therefore, quantitative results demonstrate that our method performs better in fusion performance than others.

Table 2. Validity of different modules in our method. **D** represents densely dilated feature extraction module, **A** represents the edge-guided attention mechanism and **S** expressed as skip connection.

Method	MI	CC	VIF	$Q^{AB/F}$	SCD	SSIM
Ours w/o **D**	2.65	0.59	0.34	0.34	1.61	0.35
Ours w/o **A**	2.42	**0.67**	0.44	0.40	1.50	0.41
Ours w/o **S**	2.28	0.65	0.42	0.41	1.44	0.42
Ours	**2.83**	0.65	**0.49**	**0.53**	**1.65**	**0.45**

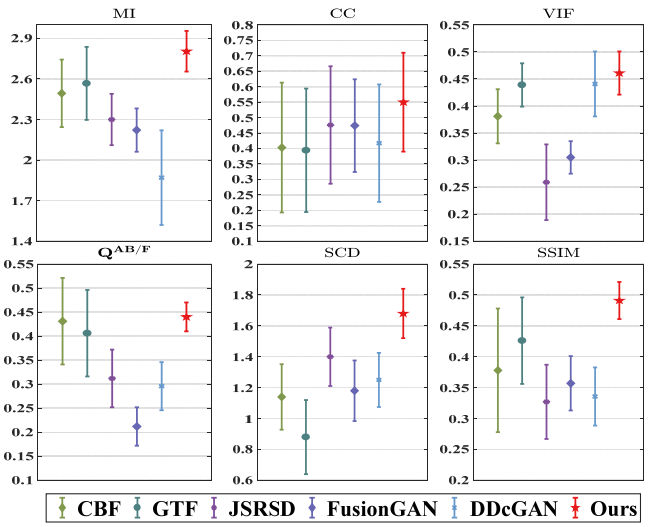

Fig. 8. Quantitative comparison of six metrics on 25 TNO image pairs. The nodes represent average value and the length of lines means variance.

3.3. Ablation Study

To verify the validity of each module in our method, quantitative ablation study is proved in Tab. 2. As shown, each module performs positively on the fusion result. Although metric CC achieves the highest score without module **A**, the other five metrics all decline significantly without it. This phenomenon lies in that module **A** calculates the salient details, while it may lose some relevant information in the fused processing. Apart from that, both module **D** and **S** perform multi-scale feature extraction and make compensate for information loss in deconvolution layers respectively. With these modules, our method is able to generate a satisfactory fusion result.

4. CONCLUSIONS

This paper developed a learning attention guided deep multi-scale feature ensemble for infrared and visible image fusion, where we designed a coarse-to-fine multi-scale feature extract module and a learning attention based fusion mechanism to overcome the limitation of the existing methods and generate a visual-friendly, noise-free fusion result. Extensive experiments on public dataset with five state-of-the-art methods comprehensively illustrate that our method can not only represent a abundance of detail texture but also avoid the noise and holes simultaneously.

5. REFERENCES

[1] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan, "Object detection with discriminatively trained part-based models," *IEEE TPAMI*, vol. 32, no. 9, pp. 1627–1645, 2009.

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE TPAMI*, vol. 40, no. 4, pp. 834–848, 2017.

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

[4] Han Xu, Pengwei Liang, Wei Yu, Junjun Jiang, and Jiayi Ma, "Learning a generative model for fusing infrared and visible images via conditional generative adversarial network with dual discriminators," in *IJCAI*, 2019, pp. 3954–3960.

[5] Junli Liang, Yang He, Ding Liu, and Xianju Zeng, "Image fusion using higher order singular value decomposition," *IEEE TIP*, vol. 21, no. 5, pp. 2898–2909, 2012.

[6] Minjae Kim, David K Han, and Hanseok Ko, "Joint patch clustering-based dictionary learning for multimodal image fusion," *Information Fusion*, vol. 27, pp. 198–214, 2016.

[7] Filippo Nencini, Andrea Garzelli, Stefano Baronti, and Luciano Alparone, "Remote sensing image fusion using the curvelet transform," *Information Fusion*, vol. 8, no. 2, pp. 143–156, 2007.

[8] Jinlei Ma, Zhiqiang Zhou, Bo Wang, and Hua 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.

[9] Hui Li and Xiao-Jun Wu, "Densefuse: A fusion approach to infrared and visible images," *IEEE TIP*, vol. 28, no. 5, pp. 2614–2623, 2018.

[10] B. K. Shreyamsha Kumar, "Image fusion based on pixel significance using cross bilateral filter," *Signal, Image and Video Processing*, 2015.

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

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

[13] Wesley J. Roberts, Jan A. Aardt Van, and Fethi Ahmed, "Assessment of image fusion procedures using entropy, image quality, and multispectral classification," *Journal of Applied Remote Sensing*, vol. 2, no. 1, pp. 1–28, 2008.

[14] Manjusha Deshmukh and Udhav Bhosale, "Image fusion and image quality assessment of fused images," *International Journal of Image Processing (IJIP)*, vol. 4, no. 5, pp. 484, 2010.

[15] Yu Han, Yunze Cai, Yin Cao, and Xiaoming Xu, "A new image fusion performance metric based on visual information fidelity," *Information Fusion*, vol. 14, no. 2, pp. 127–135, 2013.

[16] CS Xydeas, , and Vladimir Petrovic, "Objective image fusion performance measure," *Electronics letters*, vol. 36, no. 4, pp. 308–309, 2000.

[17] V Aslantas and E Bendes, "A new image quality metric for image fusion: the sum of the correlations of differences," *Aeu-international Journal of electronics and communications*, vol. 69, no. 12, pp. 1890–1896, 2015.

[18] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al., "Image quality assessment: from error visibility to structural similarity," *IEEE TIP*, vol. 13, no. 4, pp. 600–612, 2004.