Manuscript No. NCAA-D-20-01362

Original manuscript title: SESF-Fuse: An Unsupervised Deep Model for Multi-Focus Image Fusion

Neural Computing and Applications

Dear Associate Editor,

On behalf of all the authors, I would like to thank you and the reviewers for the effort put into our manuscript. We found the suggestions and comments to be very constructive and have revised the paper accordingly. We are now resubmitting a marked-up manuscript file with the new or revised text colored in red. Besides, we are providing a detailed point-to-point response in this file.

Should you have any questions, please feel free to contact me. Thank you very much in advance for your consideration of our manuscript, and we look forward to hearing from you.

Yours Sincerely,

Xiaojuan Ban

Beijing Advanced Innovation Center for Materials Genome Engineering

Responses to the comments

Reviewer #1:

The paper fuses two multi focus images.

Q1: In the related works, some of the most recently published papers that are directly related and similar to their current work is not introduced, for example the reference [6]. The reference [6] is very close to their work and should be described more, and the difference between their work and the work in [6] should be highlighted.

Answer: Thank you for your advice.

Article [6] proposed an unsupervised model for multi-focus image fusion based on gradient and connected regions, termed as GCF. We have described this paper more exhaustively in the revised manuscript, which is shown below:

HanXu proposed a gradient and connected regions based fusion method (GCF) [6], which used an encoder-decoder structure to output a gradient relation map to generate a decision map, using a deep learning model to implicitly calculate gradient information. In contrast, we use an encoder-decoder to output the original image, and spatial frequency to explicitly calculate the gradient information in deep features to generate a decision map. We speculate that our method is more robust than implicit decoder mapping.

Some other DNN based learning methods are also added in the revised manuscript.

Besides, thanks for agreement of article [6], it has cited our pre-printed article (arXiv) and compared with our method in its experimental section, which is shown in Figure 1-1. SESF-Fuse has achieved the highest results in SF and VIF evaluation metrics and lowest runtime compared to other methods (including GCF).

TABLE 1. **Mean values of different methods on metrics FMI, SF and SD on the rest 10 image pairs from the dataset. Red indicates the best, blue denotes the second best and cyan represents the third best.**

	DCTVar	DSIFT	MST_SR	GBM	CNN	DenseFuse	SESF-Fuse	GCF
FMI	0.8841	0.8900	0.8862	0.8776	0.8896	0.8598	0.8894	0.8897
SF	0.0812	0.0813	0.0808	0.0732	0.0805	0.0492	0.0815	0.0813
SD	0.2201	0.2218	0.2209	0.2035	0.2214	0.2482	0.2222	0.2219
PSNR	74.6324	74.6645	74.8515	68.6962	74.7557	71.4529	74.6482	74.6654
VIF	1.0876	1.1336	1.1204	1.0411	1.1236	0.9797	1.1407	1.1336

TABLE 2. **Average runtime comparison of different methods on the ten ten image pairs(unit: second).**

Method	DCTVar	DSIFT	MST_SR	GBM	CNN	DenseFuse	SESF-Fuse	GCF
Mean	0.712	5.037	0.585	1.250	122.095	0.536	0.312	0.654
STD	0.027	1.405	0.031	0.052	2.171	0.021	0.005	0.045

Figure 1-1. The comparison results in article [6]. Our method SESF-Fuse has achieved the best results in SF and VIF metrics and lowest runtime.

Q2: In the first paragraph in section 3.2 it is stated that "after the parameters of model are fixed". There are no explanations about these parameters and how they are fixed.

Answer: Thank you for your advice. We are sorry that this causes misunderstanding of you about our work.

There are two types of parameters in convolutional neural network. First, pre-defined parameters, which means the parameters should be fixed before training the network. Such as the depth of neural network, the size of convolutional kernel and the training settings. All these parameters are described in the manuscript. Second, trainable parameters, which means that these parameters will be optimized during the training procedure of model. For example the parameters in convolutional kernel. Machine Learning algorithms learn the pattern in data and update the trainable parameters of model in training phase. And the trainable parameters of the model will be fixed after training. Then this model with fixed parameters perform inference for new data in inference phase.

According to your suggestion, we have modified the sentence “After the training of the network has done and the trainable parameters of it has fixed,

we use spatial frequency to calculate activity level of deep feature drawn from encode in inference phase."

Q3: The description of the method in section 3.2 is hard to follow because the statements are not clearly stated.

Answer: Thank you for your advice.

According to your suggestion, we have carefully revised the section 3.2 in manuscript. Follow the revised statements of manuscript, we hope you would understand our method.

Q4: No clear algorithm is presented to describe the method. There is not enough information for the reader to implement the method and make the experiments.

Answer: Thank you for your advice.

According to your suggestion, we have carefully revised the method section in manuscript.

Besides, we have provided publicly code (https://github.com/Keep-Passion/SESF-Fuse) in manuscript, which include the trained parameters, codes and data of our method in master branch and comprehensive experiments in experiment branch. Readers can easily use this code for their own experiment.

As far as we known, many articles have already cited our pre-printed manuscript and implemented this code to conduct comprehensive experiments in their works [1, 2]. And article [6] in Q1.

With the supported information provided in the answers and related papers, we hope you would be convinced that our method is easy to implement and make the experiment for readers.

Q5: There are only some general statements for description of the method. For example, no descriptions about the parameters of the method and the kernels is presented. The statements in explanation of the proposed method are ambiguous so that it is hard to understand the meaning of the authors.

Answer: Thank you for your advice.

According to your suggestion, we have carefully revised the method section in manuscript.

Q6: The parameters of equation 4 are not describes. It is not described that how the equation is used in the model. How the equation describes the similarity?

Answer: Thank you for your advice.

Structural similarity (SSIM) is the most famous basic common sense in digital image processing [9], which uses to reflect the similarity of two images.

$$SSIM(x,y) = [l(x,y)]^{\alpha} \cdot [c(x,y)]^{\beta} \cdot [s(x,y)]^{\gamma} \quad (1\text{-}1)$$

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \quad (1\text{-}2)$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \quad (1\text{-}3)$$

$$s(x,y) = \frac{2\sigma_{x,y} + C_3}{\sigma_x\sigma_y + C_3} \quad (1\text{-}4)$$

Where x and y are two images. μ_x and μ_y are mean values of them. σ_x and σ_y are standard values. $\sigma_x\sigma_y$ is covariance of x and y. C_1, C_2 and C_3 are three constants. Generally, $C_3 = \frac{C_2}{2}$ and $\alpha = \beta = \gamma = 1$ .. $l(x,y)$, $c(x,y)$ and $s(x,y)$ are comparison between luminance, contrast and structure of x and y respectively. For example, if $\mu_x = \mu_y$, then $l(x,y) = 1$.

We modified the description in this part. And same with [1, 4], we use that to calculate the loss function in our model.

Consider the limit size of paper, we omit detail description of SSIM, but recommend the reader to reference the work [9], which was also done in many other studies in this field [1, 4].

Q7: The paper basically introduces a fusion method but, only one paragraph at the end of section 3 is used to describe the fusion. Equation 9 also is not described.

Answer: Thank you for your advice. The key ingredient in multi-focus image fusion is the calculation for decision map. The other sub-section of section 3 is used to draw out the decision map. We revised the manuscript and described the Equation 9.

Q8: In the section 4.1, the experiments are described, but, some ambiguities exist. In the first paragraph in page 5 it is stated that "And then we used acquired parameters to perform SF fusion on the testing set above." What are the parameters? How they are determined and use?

Answer: Thank you for your advice.

Same with Q2, we are sorry again that this causes misunderstanding of you about our work. The parameter in this sentence is the trainable network parameters, for example parameters in convolutional kernel.

According to your suggestion, we have modified the sentence “And then after training has done and the trainable parameters in network has fixed, we used the acquired parameters to perform image fusion on the testing set above”.

Q9: In table 1, three quality metrics are used as comparison, but no description about the difference among these is present.

Answer: Thank you for your advice. These three metrics are described in the cited literature [10]: Q_g and Q_m are image feature based metrics. Q_g is a gradient based fusion metric to evaluate the amount of edge information. Q_M is implemented with a two-level Haar wavelet to calculate edge information from the high and band-pass components of the decomposition. Q_{cb} is a human perception inspired fusion metric to reflect the visual quality of fusion result.

And thanks for your suggestion, we have described the three qualitative metrics in the revised paper to help the readers to understand our paper.

Q10: Details of the experiment results in table 2 is not present. For example, the number of train and test samples. Are there any parameters for the compared methods that should be set? If yes, what is the setting process. What are the results of the experiments on other datasets? At least one other data set is needed to be examined.

Answer: Thank you for your advice.

There is misunderstanding of structure in experiment section. That is section 4.3 and section 4.4 have same experiment setting which has described in section 4.1. We use 82783 images in MS COCO as training set and 40504

images as validation set. And we use 38 pairs of multi-focus images as testing set for evaluation.

The implementation of compared methods are drawn from publicly codes and the related parameters are set to the recommended values introduced in related publications. We have modified the corresponding part in section 4.2 in our revised paper.

Because there are few public datasets [7, 8] in the field of multi-focus image fusion, and the number of data in each dataset is relative small, we merge two datasets [7, 8] to form an uniform test set to conduct each experiment in our work, which was also done in many other studies in this field [3, 4, 5]. In addition, following the pre-print version of our article, a new dataset MFFW [2] has established and conducted comprehensive experiments with our SESF-Fuse and other methods, as shown in Figure 1-2. According to the experiment in MFFW, SESF-Fuse also achieves the highest performance on gradient based metrics, such as spatial frequency metric (SF), average gradient (AG) and gray level difference (GLD).

TABLE II. METRICS ON MFFW2. THE BEST, THE SECOND BEST AND THE THIRD BEST RESULTS ARE HIGHLIGHTED BY BOLD TYPEFACE, RED AND BLUE, RESPECTIVELY.

	BF	BRWTSFM	CBF	CSR	DSIFT	GFDF	GFF	MSTSR	MWGF	CNN	IFCNN	MMFNet	SESF
MI	1.11041	1.04145	0.86481	0.90291	**1.11394**	1.07927	0.93709	0.79136	1.05172	1.06379	0.77777	0.91739	1.06069
TE	**0.37851**	0.36902	0.35711	0.36007	0.37730	0.37460	0.36285	0.34349	0.36948	0.37433	0.34570	0.35683	0.36670
NCIE	0.83999	0.83553	0.82578	0.82669	**0.84017**	0.83772	0.82969	0.82192	0.83565	0.83685	0.82105	0.82846	0.83681
GBM	0.59408	0.61646	0.52436	0.54657	0.63253	0.63469	0.60216	0.60297	0.65209	**0.66146**	0.43159	0.42425	0.60389
SF	-0.06188	-0.04210	-0.06579	-0.05860	-0.03101	-0.04540	-0.05152	-0.03419	-0.05450	-0.05786	-0.06056	-0.04734	**0.00158**
SSBM	**0.98779**	0.94442	0.85081	0.87093	0.93205	0.97328	0.95175	0.91612	0.98216	0.97580	0.87969	0.88438	0.95831
CBM	0.74556	0.72685	0.65541	0.68420	0.72038	0.74257	0.71857	0.69597	**0.74643**	0.73620	0.64275	0.66439	0.73005
LIF	0.38648	0.38565	0.38885	0.38495	0.38700	0.38603	0.38611	0.38314	0.38634	0.38479	0.38292	**0.37826**	0.38404
AG	3.51687	3.56747	3.49238	3.45226	3.64229	3.55903	3.52980	3.64965	3.51950	3.51256	3.54638	3.60238	**3.66501**
MSD	0.07902	0.07913	0.07818	0.07895	0.07915	0.07907	0.07904	0.08046	0.07907	0.07989	**0.08054**	0.07894	0.07954
GLD	17.39289	17.63803	17.26349	17.00870	17.99025	17.59924	17.45074	18.03521	17.39076	17.36946	17.52463	17.80669	**18.10413**

Figure 1-2. The comparison result in MFFW dataset [2]. Our method SESF-Fuse has achieved the best results in SF, AG and GLD metrics compared to other methods.

Q11: The experiments also described by some ambiguous statements that is hard to follow. No pseudo code, flowchart is presented for the method and the experiments to clarify the method.

Answer: Thank you for your advice.

The network architecture is the most import part for the description of deep learning based algorithms. We have described it carefully and added flowchart in our manuscript with Figure 1 in manuscript. Because the basic implementation is same in deep learning based methods (training and testing phase), we did not add pseudo code considering the limit size of the paper length, which was also done in many other studies in this field [3, 4, 5].

We thought that there is an improper description of our method, we have revised paper according to all of your questions. We hope you can understand our method according to the responses in this file.

References:

[1] Huang J, Le Z, Ma Y, et al. A generative adversarial network with adaptive constraints for multi-focus image fusion [J]. NEURAL COMPUTING & APPLICATIONS, 2020.

[2] Xu S, Wei X, Zhang C, et al. MFFW: A new dataset for multi-focus image fusion [J]. arXiv preprint arXiv:2002.04780, 2020.

[3] Liu Y, Chen X, Peng H, et al. Multi-focus image fusion with a deep convolutional neural network [J]. Information Fusion, 2017, 36: 191-207.

[4] Li H, Wu X J. DenseFuse: A fusion approach to infrared and visible images [J]. IEEE Transactions on Image Processing, 2018, 28(5): 2614-2623.

[5] Prabhakar K R, Srikar V S, Babu R V. DeepFuse: A Deep Unsupervised Approach for Exposure Fusion with Extreme Exposure Image Pairs[C]//ICCV. 2017, 1(2): 3.

[6] Xu H, Fan F, Zhang H, et al. A Deep Model for Multi-Focus Image Fusion Based on Gradients and Connected Regions [J]. IEEE Access, 2020, 8: 26316-26327.

[7] Nejati, M., Samavi, S., Shirani, S.: Multi-focus image fusion using dictionary based sparse representation. Information Fusion. 25, 72–84 (2015).

[8] Savi´c, S., Babi´c, Z.: Multi focus image fusion based on empirical mode decomposition. In: 19th IEEE International Conference on Systems, Signals and Image Processing (IWSSIP) (2012).

[9] Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE transactions on image processing, 2004, 13(4): 600-612.

[10] Liu Z , Blasch E , Xue Z , et al. Objective Assessment of Multiresolution Image Fusion Algorithms for Context Enhancement in Night Vision: A Comparative Study[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2012, 34(1):94-109.

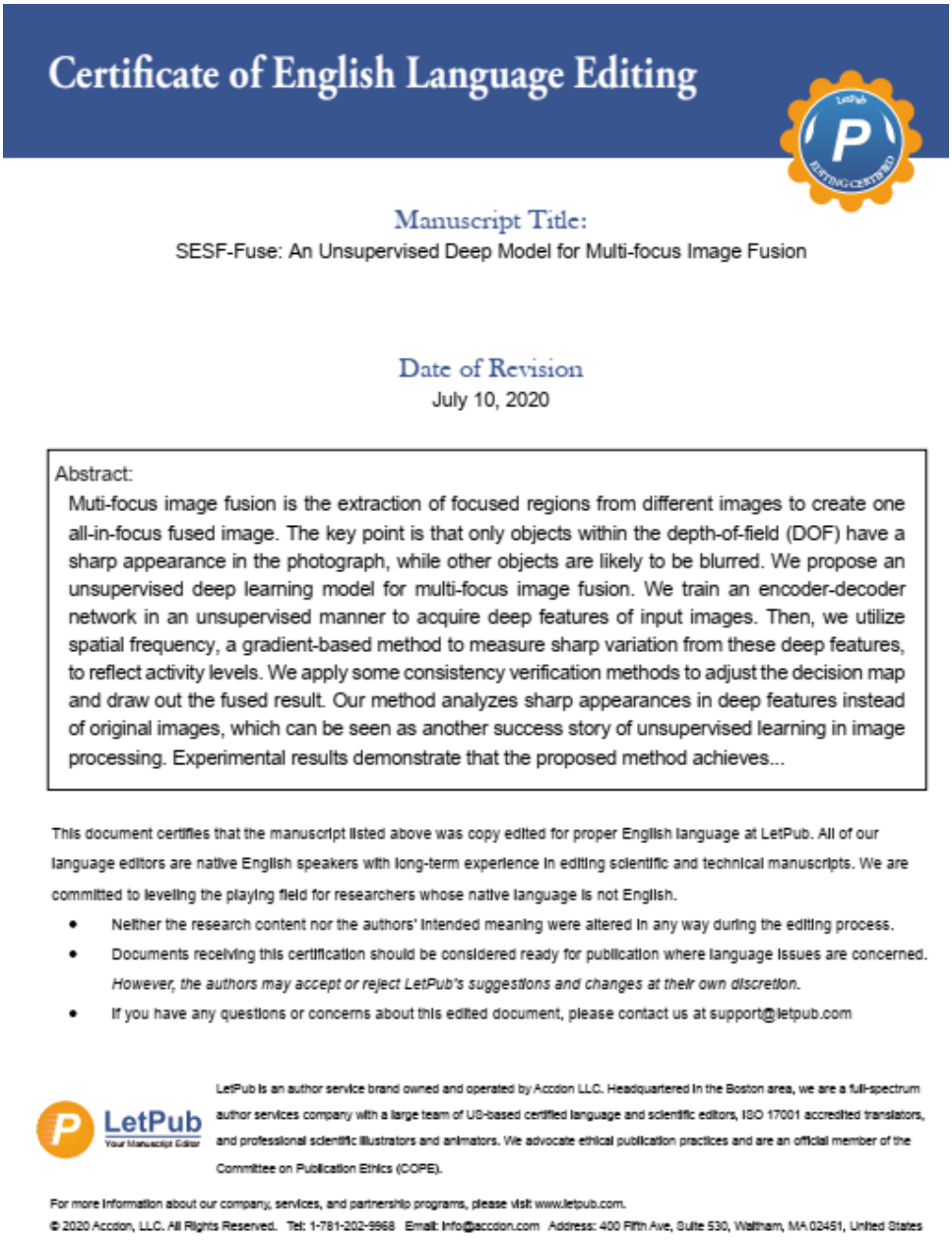
Reviewer #2:

In this paper the authors present an unsupervised deep model for fusing multi-focus images, which is a problem worth researching. The paper reads well and it is well organized. However, there are a series of (mainly major) issues:

Q1: I have seen several English issues, particularly regarding grammar. Please, ask a native speaker or proofreading service to revise the English of this paper.

Answer: Thank you for your advice.

We have asked English proofreading service provided by LetPub to proofread the entire paper. The certificate is shown below:



Certificate of English Language Editing

Manuscript Title:
SESF-Fuse: An Unsupervised Deep Model for Multi-focus Image Fusion

Date of Revision
July 10, 2020

Abstract:
Muti-focus image fusion is the extraction of focused regions from different images to create one all-in-focus fused image. The key point is that only objects within the depth-of-field (DOF) have a sharp appearance in the photograph, while other objects are likely to be blurred. We propose an unsupervised deep learning model for multi-focus image fusion. We train an encoder-decoder network in an unsupervised manner to acquire deep features of input images. Then, we utilize spatial frequency, a gradient-based method to measure sharp variation from these deep features, to reflect activity levels. We apply some consistency verification methods to adjust the decision map and draw out the fused result. Our method analyzes sharp appearances in deep features instead of original images, which can be seen as another success story of unsupervised learning in image processing. Experimental results demonstrate that the proposed method achieves...

This document certifies that the manuscript listed above was copy edited for proper English language at LetPub. All of our language editors are native English speakers with long-term experience in editing scientific and technical manuscripts. We are committed to leveling the playing field for researchers whose native language is not English.

- Neither the research content nor the authors' intended meaning were altered in any way during the editing process.
- Documents receiving this certification should be considered ready for publication where language issues are concerned. *However, the authors may accept or reject LetPub's suggestions and changes at their own discretion.*
- If you have any questions or concerns about this edited document, please contact us at support@letpub.com

LetPub
Your Manuscript Editor

LetPub is an author service brand owned and operated by Accdon LLC. Headquartered in the Boston area, we are a full-spectrum author services company with a large team of US-based certified language and scientific editors, ISO 17001 accredited translators, and professional scientific illustrators and animators. We advocate ethical publication practices and are an official member of the Committee on Publication Ethics (COPE).

For more information about our company, services, and partnership programs, please visit www.letpub.com.
 Tel: 1-781-202-9968 Email: info@accdon.com Address: 400 Fifth Ave, Suite 530, Waltham, MA 02451, United States

Figure 2-1. The certificate of English Language Editing

Q2: In the abstract and the Introduction, the authors state that their method "achieves" the state-of-the-art. What do you mean by that? Does your method outperforms the s-o-t-a methods? Is it as good as they are? In all cases?

Answer: Thank you for your advice.

It is not rigorous academic representation for ‘sota’. According to experiment in our paper and re-implemented experiment in other works [1,2,3],

our method explicitly calculate gradient information in inference phase and use it to generate decision map, so that it is incline to preserve the clarity region which have more gradient information and shows the SOTA performance in gradient based metrics compared to other methods. However, there is other problem in multi-focus image fusion, such as defocus spread effect, which cannot be handled with SESF-Fuse method, which we will described more detailed in Q5.

Thank you for your suggestion again, we would change the description that：Our method achieves state-of-the-art in the gradient based fusion evaluation compared to other methods.

Q3: Please, provide more details about section 3.3.3 and explain the formula in a more comprehensive way.

Answer: Thank you for your suggestion. We have provided more detail in section 3.3.3 and revised the formula comprehensively.

Q4: You should provide more details about indicators Qg, Qm and Qcb in section 4.2. Please, explain those indicators although they are described in the cited literature so that readers do not have to consult those references to understand your paper.

Answer: Thank you for your suggestion. We have added the description of those three metrics in our revised paper to help the reader to understand our work.

Q5: Discussion section needs to be widen. Please, discuss the results in a deeper way. Indicate some more future lines of research and provide more details about them.

Answer: Thank you for your suggestion. We have widen the discussion to describe the limitation of SESF-Fuse and described more future lines of research, which is shown below:

First, SESF-Fuse cannot fuse multi images in an efficient way. Most of currently methods inefficiently fuse images one by one in series. The main reason behind it is that the output image from decoder cannot exactly restore the real pixel value, which causes a low Qg value in evaluation. So that these methods try to acquire an intermediary result (decision map) to recover true pixel value of source image. And some post-processing methods can only be applied to decision map in two images fusion manner, which obstacles the efficient multi image fusion. Therefore, the ability of decoder of fusion network should improve to solve the problem. If decoder can recover the source pixel value precisely without any post processing method, the network will perform feature extraction by multiple encoders and fuse them by one decoder to save inference time. GAN and well-designed loss function may be a good solution to address this problem.

Second, defocus spread effect, which described in [4] and shown in Figure 2-2, cannot be addressed by SESF-Fuse. Article [4] has proposed a novel deep learning model, MMFNet, which employ new data generation method and loss function to handle this problem, and produce better fusion result in objective assessment with defocus spread effect, shown in Figure 2-3. However, according to the re-experiment of [1], MMFNet achieves the lower performance compared to SESF-Fuse in gradient based metrics. We speculated that MMFNet acts like guided filter to smooth the boundary of focus and defocus regions to produce better visualization fusing result. However, it will decrease the gradient based subjective evaluation result due to smooth operation. We thought that there will be a better way to combine the advantages of two methods.

(a) Foreground focus (b) Background focus

Figure 2-2. Different defocus spread effects when the foreground or the background is out of focus, shown in enlarged real-world images. The in-focus boundary of foreground object is labeled red. When the foreground object is in focus, but the background is defocused, the foreground object will not be influenced by the defocus of the background. In contrast, when the foreground object is defocused, the defocus spread effect will lead to a blurry object bigger than the original focused object [4].

MMFNet SESF-Fuse

Figure 2-3. A demonstration of fusing result influenced by defocus spread effects. [1]

Multi image fusion and defocus spread effect may appear in many industrial applications, for example imaging of electronic components. We thought there are many improvements should be conducted in future work.

Q6: Another minor issue: In section 3.3.1, in the paragraph previous to formula (5), please indicate the abbreviation SF this way: "We calculate its Spatial Frequency (SF) using ..."

Answer: Thank you for your advice. We have modified this sentence.

References:

[1] Xu S, Wei X, Zhang C, et al. MFFW: A new dataset for multi-focus image fusion [J]. arXiv preprint arXiv:2002.04780, 2020.

[2] Xu H, Fan F, Zhang H, et al. A Deep Model for Multi-Focus Image Fusion Based on Gradients and Connected Regions [J]. IEEE Access, 2020, 8: 26316-26327.

[3] Huang J, Le Z, Ma Y, et al. A generative adversarial network with adaptive constraints for multi-focus image fusion [J]. NEURAL COMPUTING & APPLICATIONS, 2020.

[4] Ma H , Liao Q , Zhang J , et al. An {\alpha-Matte Boundary Defocus Model Based Cascaded Network for Multi-focus Image Fusion [J]. 2019.

Reviewer #3

It is missing in the notice email which send to us.

Reviewer #4:

This paper targets the multi-focus image fusion issue. Multi-focus image fusion needs to extract the focused regions from the different images and fuse them into one image. The previous works leverage the norm of feature vectors to represent the activity level in an image. However, this method can only be applied to fuse infrared and visible images while it is not suitable for a color image. In this paper, the authors assumed that only the objects within the depth-of-filed have sharp appearances in the photograph while other objects are likely to be blurred. Therefore, they proposed a fusion method based on an unsupervised deep convolutional network. It first leverages the encoder-decoder to extract deep features. Then it calculates the activity level using those deep features in the inference phase. Finally, it obtains the decision map to fuse two multi-focus source images.

Strength:

1) This paper has a strong motivation. It first clearly lists the key feature in multi-focus image fusion is that only the objects within the depth of field have the sharp appearances. In order to leverage this feature, it presents a fusion method based on unsupervised deep convolutional network and spatial frequency technique.

2) Although the spatial frequency method is intuitive and simple, the further verification method in this paper is novel. It applies the morphology operator as a post-process to rectify the decision map. It also gives the specific parameter values chosen from preliminary experiments.

3) The experiment in this paper is comprehensive. First, the experimental setup is clearly demonstrated. 16 representative image fusion methods are compared with the proposed method in this paper, which is solid. Moreover, this paper leveraged three popular metrics to evaluate fusion quality.

Answer: Thank you for your agreement of our work, we will response to your questions one-by-one as follow:

Weakness and Questions:

Q1: My first concern is there are only several related works are discussed and compared in the introduction and related work parts. As far as know, there are a lot of DNN-based multi-focus image fusion works, it would be better if the authors can have a more comprehensive and clear Pros and Cons analysis of these works.

Answer: Thank you for your advice. We have added some DNN-based multi-focus image fusion works and analyzed some of them in related work of manuscript.

Q2: Although the authors mentioned that the three SE modules design is inspired by another paper (unfortunately, [22] lacks a paper name in the inference page), I still confused about the motivation of such design. I recommend the authors add a more specific description to explain why including SE modules in the encoder-decoder. Moreover, in DenseFuse, it leverages encoder to extract useful deep feature while using the decoder to reconstruct the aiming image. I wonder why this paper combines both encoder and decoder to obtain the deep features.

Answer: Thank you for your advice.

We found that there is mistake in latex coding with [22] and have been modified in the revised version.

And according to your suggestion, we have added a description of SE module in our revised paper, which is shown below:

“Recently, squeeze and excitation (SE) module achieves promising performance on image recognition and image segmentation, owing to it can effectively enhance spatial feature encoding by adaptively re-calibrating channel-wise or spatial feature responses. Therefore, we introduce SE module into our network in order to enhance the robustness and representativeness of deep features.”

The difference with SESF-Fuse and DenseFuse：

DenseFuse leverages encoder to extract useful deep feature while using the decoder to reconstruct the aiming image. It uses decoder in both training and inference phase. By Contrast, different with DenseFuse, we only use decoder in training phase to help training a robust encoder. And in inference phase, we use encoder to exact deep features and then use spatial frequency to calculate activity level.

We have conducted an experiment in ablation study (section 4.3) and re-written it to make it more easily to understand:

Table 4-1 shows mean average score with different methods. 'sf' refers to spatial frequency. 'dm' means the algorithm use decision map which export from encoder to fuse images. For the algorithm without 'dm' term, it denotes that the network directly output fusing result from decoder, which receives the input of fusion feature. DenseFuse compared ‘add’ and ‘l1-norm’ fusion strategies and drew out the conclusion that l1-norm of deep feature could be used to fuse infrared-visible images. They utilized feature intensity to calculate activity level. We found that cSE_l1_norm achieve smaller performance in three gradient based quality metrics compared to cSE_sf. Thus we speculated that feature gradient (calculated by spatial frequency) is suited to multi-focus fusion task.

In addition, even though the deep learning has promising representative ability, it cannot recover the image perfectly. Thus if we use sf to fuse the deep features and input to decoder, then draw out result (cSE_sf), the fused result can not completely recover every detail of in-focus region. And the performance values in three metrics are relatively low. Therefore, we propose to use deep features drawn from encoder to calculate the decision map and use it to generate the fused result (cSE_sf_dm). This operation will precisely recover the real pixel value of source image, and the performance of cSE_sf_dm is higher than cSE_sf.

Table 4-1. Experiments with different settings

Methods	Q_g	Q_m	Q_{cb}
cSE_l1_norm	0.5124	2.4961	0.6020
cSE_sf	0.6885	2.7216	0.7526
cSE_sf_dm (SESF-Fuse)	**0.7105**	**2.8886**	**0.7848**

Q3: In this paper, the images in the training set are transformed into gray-scale. I am curious that what' the difference between the infrared image and the gray-scale image. Because the paper mentioned that the L1 norm method can be only applied to infrared images while the spatial frequency can be applied to the multi-focus task.

Answer: Thank you for your advice.

The images in the training set are transformed into gray-scale is to reduce the parameters in the network.

Beside, we speculated that there is a difference between multi-focus image fusion and infrared-visible image fusion. As shown in the first row of Figure 4-1, the difference in source 1 and source 2 in multi-focus image fusion is the focused region. Generally, the focused region have clarity information which is shaper than defocus region. That is why spatial frequency, which is a gradient based indicator can be used to detect focused region. And in our work, we have demonstrated that spatial frequency achieves higher performance compared to l1-norm.

By contrast, the second row is a demonstration of infrared-visible fusion (driven from DenseFuse). The key difference is that the objects with thermal radiation will have more intensity in infrared image compared to visible image. For example, the person in the first column and second row. And the red rectangle part denotes that the light from the bulb is cold light without thermal radiation, that is can be seen in visible image (source 2) and not be seen in infrared image (source 1). So that the key difference is the pixel intensity not the pixel gradient. That’s why we thought l1-norm is much suited to infrared-visible image fusion. Besides, if the object with thermal radiation is very large, this region will be totally white in infrared image and be totally black in visible image, which will lead to same gradient in both of them. So that, we speculated

that gradient based indicator, spatial frequency for example, cannot be used in infrared-visible image fusion which will causes detection error.

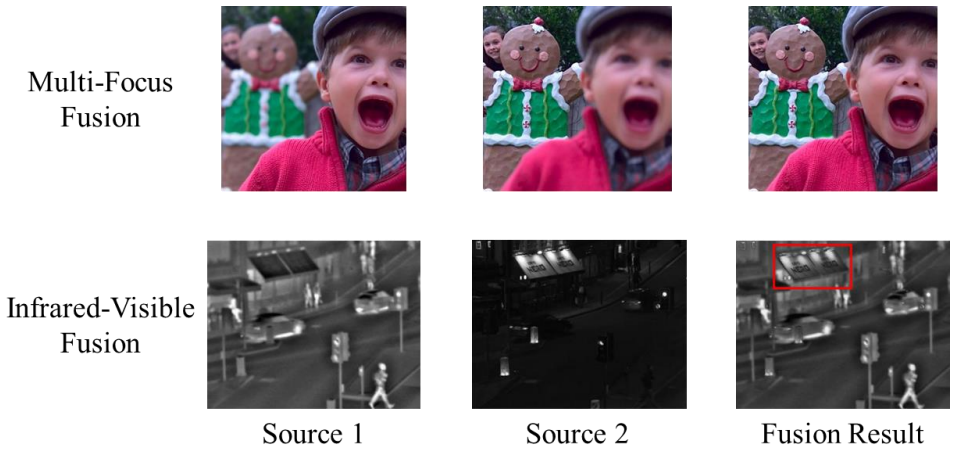


Figure 4-1. The demonstration of multi-focus fusion and infrared-visible fusion

Q4: The reference format is bad. A lot of references lack paper names, such as [3], [22], [24].

Answer：Thank you for your advice. We are so sorry for the mistake in latex coding, which missing all the titles in the reference section. We assumed that this might be a bug in official template with statement '\bibliographystyle{spphys}'. We have checked the latex coding and revised this mistake. Note that we have changed the temple of references, so the order of it has also changed.

Overall, this paper proposes a multi-focus image fusion method which is based on deep convolutional network and spatial frequency technique. This paper has strong motivation and clear methodology design. The experiment is very comprehensive and sufficient SOTA methods are compared. However, in the related work part, a few works are mentioned and discussed. Also, the motivation of use three SE modules is not clear.

Reviewer #5:

In this manuscript the authors have proposed a novel unsupervised deep learning model to handle multi-focus image fusion task. This task extracts focused regions from different images to create one all-in-focus fused image. The authors have presented unsupervised deep learning model, that is a type of machine learning, to address multi-focus image fusion problem. This model sheds lights on the potential of combination of deep learning and traditional vision algorithm in multi-focus fusion exploring its advantages over previous fusion methods.

The manuscript is well written, scientifically sound and contains interesting results to be considered for publication in the Neural Computing Application journal. However, before the acceptance the following points should be addressed by the authors:

Q1: The first point concerns the proposed model and some related questions: i) how the fluctuations always present in all systems are taken into account? I mean, what about the fluctuations in the near focused image and far focused image? The authors should comment deeply on this point.

Answer: Thank you for your advice.

We agree with your statement. Although the decision map calculated from deep features is robust enough to recognize the location of focused region, there may be some small faults by inappropriate decisions caused by fluctuations, or noises, appears in image generated from imaging system. For example, there may have gaussian or salt-and-pepper noises randomly occur on each pixel during imaging, which influence the calculation of activity level. Especially, salt-and-pepper noise will seriously impact the calculation result of spatial frequency. Therefore, we propose to use some consistency verification methods to handle it. Morphology operations, such as opening and closing, and

small region removal strategy will eliminate the incorrect focused pixel in decision map.

We have modified this in section 3.3.3.

Q2: The second point concerns the technique of the model, I guess that in the fusion model some fluctuations are implicitly taken into account. This seems to be similar to the benefit of noise in visual perception due to the stochastic resonance phenomenon. See for example the following papers: i) Riani, M. & Simonotto, E. Stochastic resonance in the perceptual interpretation of ambiguous figures: A neural network model, Phys. Rev. Lett. 72, 3120–3123 (1994); ii) Simonotto, E. et al. Visual perception of stochastic resonance. Phys. Rev. Lett. 78, 1186–1189 (1997); iii) Elena Itzcovich, Massimo Riani & Walter G. Sannita, Stochastic resonance improves vision in the severely impaired, Scientific Reports 7, 12840 (2017) and references therein. The authors should compare their model with that one proposed in all the abovementioned papers, in which the stochastic resonance phenomenon plays a fundamental role in visual perception.

Answer: Thank you for your advice.

Stochastic resonance (SR) is a phenomenon resulting from the effect of a random or unpredictable interference (“noise”) on information processing in nonlinear threshold systems. Noise added in optimal amounts enhances the information transfer and improves the detection of sub-threshold signals. As described above, there may be some noises appear on the image and influence, and this may generate SR phenomenon. We speculated that it will decrease the impact of noises in the calculation of activity level.

Thank you for your suggestion, we have compared and cited these papers in manuscript.

Q3: Due to the importance of noise in the visual perception, the authors should insert a special paragraph in the revised version of the manuscript dedicated to the positive role of noise in physical and biological systems and stochastic resonance phenomenon in particular. Here some suggested references for this new paragraph: i) Entropy 19, 20 (2017); ii) Chaos, Solitons & Fractals 81, 412-424 (2015); iii) Eur. Phys. Journal B 65, 361-367 (2008); iv) Phys. Rev. B 91, 235412 (2015); v) Phys. Rev. B 87, 214515 (2013); vi) Ecol. Model. 213, 449-462 (2008); vii) Acta Physica Pol. B 35, 1491-1500 (2004).

Answer: Thank you for your advice.

We have carefully read the papers you have mentioned. Although it was a hard work to understand these papers for beginners like us, we speculated that noise in physical and biological system have positive role in the field of image fusion. For example, if we added the external fluctuations to intrinsically noisy systems in training or inference phase, it will induce an enhancement of the dynamical stability of deep learning model, resulting in a less noisy response and robust feature encoder.

In such short revisiting time, we have inserted a paragraph and tried our best to discuss the positive role of noise in physical and biological systems and stochastic resonance phenomenon. And we have cited some of these works in our manuscript.

We will conduct experiment to investigate the positive role of noise for image fusion in another work.

Q4: Finally, a minor point, the authors should explicitly write all the acronyms the first time they used them in the manuscript, for example: i) convolutional neural network (CNN); ii) standard L1?; iii) GAN.

Answer: Thank you for your advice. We have modified it in the first time we used them.