Image Fusion with Deep Leaning using Wavelet Transformation

Akansha Sharma Hitakshi Gupta Yashika Sharma
Student Student Student
Department of Computer Science
Banasthali University, Tonk, India

Abstract: The method of combining significant details from two or more source images of a scene into a final fused image is known as Image Fusion. When compared to any of the other input images, the fused output image will have more detailed information in it. The objective of image fusion is to obtain the most desirable data from each image. Elementary requirement for image fusion is that fusion process should comprise all appropriate information with concentrated noise and it should not introduce any artifact in the fused image.
The objective of this work deals with the edge detection approach for multi-focused images by means of complex wavelets-based image fusion. Many of the existing fusion algorithms extract the high-frequency and low-frequency information by designing some filters and then adopt different fusion rules to obtain the fused image. This study presumes that Discrete Wavelet Transform is one of the best and most effective algorithms for Image Fusion. So a wavelet is used for the multiscale decomposition of the source and fused images to obtain high-frequency and low-frequency images. To acquire more clear and complete fused image as result, deep convolutional neural network is used to learn the direct mapping between high-frequency and low-frequency images of the source and fused images.

Index Terms **– Image Fusion, Deep Learning, Wavelet Transform, CNN, DWT, Multi-focus image.**

Introduction

Image fusion is the only possible way to repossess corresponding information from different imaging modalities. Elementary requirement for image fusion is that fusion process should comprise all appropriate information with concentrated noise and it should not introduce any artifact in the fused image. Pixel level fusion, Feature level fusion and Decision level fusion are the three major types of image fusion. At the pixel level, the pixel of the first image is already registered and related to the second image in the database and then the same pixel from the second image is analyzed. At the feature level, the feature of any object in the image is matched with the object from another image and then the features are fused to get a new better image. And at the decision level, both the images are analyzed separately and the information regarding each of the image, say feature and characteristics, are stored and then that collected information is fused to get a new complete and fused image. Most of the image fusion algorithms are focused on pixel level image fusion as it is simple and computationally efficient. Pixel level fusion techniques are variable from simple spatial domain techniques to transform domain techniques.

The key problem of image fusion is how to extract the salient features from the source images and how to combine them to generate the fused image. For decades, many signal processing methods have been applied in the image fusion field to extract image features, such as discrete wavelet transform(DWT), contourlet transform, shift-invariant shearlet transform and quaternion wavelet transform etc. [1].

Currently deep learning has gained many breakthroughs in various computer vision and image processing problems, such as classification, segmentation, super-resolution and many more. In the field of image fusion, the study based on deep learning has also become a very active topic in last few years. A variety of deep learning based image fusion methods have been proposed for digital photography (e.g. multi-focus image fusion, multi-exposure image fusion), multi-modality imaging (e.g. medical image fusion, infrared/visible image fusion) [2]. The realization of multi-focus image fusion is of practical significance. The focus range of the visible light imaging system on the target area is limited by the depth of field of the optical system. In an image that is generated for the same scene, only the vicinity of the focus is clear and other objects are blurred to varying degrees. Multifocus image fusion technology can fuse differently focused images to generate a single image and combine some objects or information to obtain a more accurate description. Multifocus image fusion can overcome the limitations of a single sensor in terms of spatial resolution, geometry, and spectrum to improve the reliability of image processing [3], such as through feature extraction, object recognition, edge detection and image segmentation. Multifocus image fusion technology has been widely used in remote sensing, medical imaging, military, transportation and machine vision. The fused image obtained by the image fusion method based on the transformation domain is usually accompanied by image distortion and other phenomena. Therefore, determining a new multi-focus image fusion algorithm has important theoretical significance and practical value. The key to multi-focus image fusion is to extract the information of the clear part of the two images for fusion processing. In this work the deep learning method is used to learn the direct mapping between the source image and the fused image. The deep convolutional neural network is used to train the clear image and its corresponding blurred image to encode the mapping. The fusion rules of multi-focus images can be generated through CNN model learning. On the basis of this idea, wavelet transform is also used in this work to extract the high-frequency and low-frequency information of the image and inversely transforms the fused high-frequency and low-frequency information into the fused image. The low-frequency sub-band of the image contains the key features of the image and the high-frequency sub-band of the image contains the detailed information of the image, which is related to the sharpness of the image. A convolutional neural network is used to learn the direct mapping between the high-frequency and low-frequency sub-bands of the source of the source and fused images, respectively [4] and obtain the fusion rules of the low-frequency and high-frequency sub-bands. These rules determine the low-frequency and high-frequency information of the fused image [5]. For supervised image classification, a conventional CNN consists of: 1) Convolutional layers for feature/representation learning, which utilize local connections and shared weights of the convolutional kernels followed by pooling operators, resulting in translation invariant features and 2) fully connected layers for classification, which use high-level

image features extracted from the convolutional layers as input to learn the complex mapping between image features and labels. CNN is a suitable platform to test and compare the different fusion strategies as we can customize the fusion location in the network structure: either at the convolutional layers, fully connected layers, or network output [6].

I. RELATED WORK

Image fusion can be classified into three levels: pixel level fusion, feature level fusion and decision level fusion. The pixel level fusion involves comprehensive processing using the pixel values of the image. The pixel level based methods of image fusion can be categorized in two categories as spatial domain based image fusion methods and transformation based image fusion methods. The image fusion methods based on the spatial domain includes selecting the pixels of clearly visible parts of the images to form a completely visible fused image. Transformation based image fusion methods generally decompose the original image into different coefficients and then they fuse these transformation coefficients by the corresponding fusion rules and finally acquire the fused image by reconstruction of the fusion coefficients. The spatial domain based approach has the advantage of directly fusing the focal region of the source image; however, this method of image fusion is highly dependent on the choice of clear measurement criteria, such as gradient energy, spatial frequency or standard deviation of the image. Since the structure information cannot be represented by a single pixel, the spatial domain based fusion methods require efficient extraction of the focus area from the source image [5].
Second, the feature level based methods depend on synthetic features and structural characteristics of images, such as edges, corner points, and textures to segment the image or get a target distribution information from a local area of image. Then, information from the source images will be extracted and combined by applying certain fusion rules. The representative methods are based on object detection, edge extraction, image segmentation etc. The feature level based image fusion methods require a manual feature selection, as well as a manually designed fusion rule, and the fusion performance very much depends on the features and fusion rules.
Last, the decision level fusion is the most advanced option among the three levels of image fusion, in this a decision is made to incorporate the target based on a discriminative information according to a designed fusion rule. The fusion strategy is based on learning-based classifiers that generally quantify the reliability of classification. The limitation of decision level image fusion method is the high dependency on detection of classification results [7].

The existing multi-focus image fusion algorithms, precisely, the image fusion algorithms based on the spatial domain, focus on proposing a new model, planning more complex fusion rules, or finding an index to measure the resolution of image pixels or sub-bocks for guiding image fusion. However, a single image feature cannot be applied suitably to a variability of composite image environments, and it is practically impossible to design an ideal fusion model that considers all factors.

Liu et al. [8] used a deep neural network for multi-focus image fusion, the network that is used is fundamentally a classification network, which may direct to an imprecise boundary between the focused and unfocused regions of the image. Xu et al. [34] attempted to use images with different focus for end-to-end mapping and establish many-to-one mapping between the source and output images, a full convolutional dual-stream network architecture was designed to realize pixel-level image fusion. Zhao et al. [10] proposed the use of a multilevel deep supervised convolutional neural network for multi-focus image fusion and the design of an end-to-end network, through which joint generation feature extraction, fusion rules, and image fusion could be learned. Zhao et al. [10] constructed a new model to fuse the captured low frequency features with high frequency features, the features of wavelet multiscale transformation were also used to decompose the image in low frequency and high frequency domains. Minguri et al. [11] developed a pixel-by-pixel convolutional neural network to recognize the focus and defocus pixels in the source image for multi-focus image fusion according to the neighborhood information.

In past few year, machine learning algorithms have been widely used in various kind of image fusion, and also achieved success in the image fusion field. At first, Yang et al. [12] stated the sparse representation technique to fuse the multi-focus images, in which the image patches were represented with an over complete glossary and corresponding sparse coefficients, and then the input images were fused through fusing the sparse coefficients of each pair or set of image patches. Deep learning techniques, specifically the convolutional neural network (CNN), have took new evolution into the field of image fusion [13]. At first, Liu et al. [14] presented CNN to fuse multi-focus images. They formulated multi-focus image fusion as a classification assignment and used CNN to predict the focus map, as each pair of image patches could be classified into two categories: 1. First patch was focused and another one was blurred and 2. First patch was blurred and another one was focused. Tang et al. [15] proposed a CNN model to learn the effective focus measure (i.e., metric for quantifying the sharpness degree of an image or image patch) and then compared the focus measures of local image patch pairs of input images to determine the focus map. After that, the above two algorithm processed the focus images according to the refined focus maps. Song et al. [16] applied two CNNs to fuse the spatiotemporal satellite images, i.e., large-resolution MODIS and low-resolution land-sat images. Specifically, they respectively used two CNNs to perform super-resolution on the low-resolution land-sat images and extract image features, and then adopted high-pass modulation and weighted strategy to reconstruct the fusion image from the extracted features similar to the transform domain image fusion techniques [17]. Even though the convolutional neural network models have achieved more or less success in the image fusion field, the current models lack the simplification ability and could simply perform well on specific type of images. This problem will nevertheless bring us some trouble in developing the CNN based algorithms for fusing images without ground-truth images e.g. CT-MR images and infrared-visual images. Moreover, most of the proposed convolutional neural network models are not designed in the end-to-end manner, and thus require additional procedures to complete the image fusion. Overall, the CNN based image fusion models have not been fully exploited for the image fusion task, thus there is still much space to improve the architectures of the CNN based image fusion models, so as to increase their performance and generalization ability. Through comparing the transform domain image fusion algorithms and CNN based image generation models, we find there are several similar characteristics between these two kinds of algorithms. Primarily, the transform domain algorithms usually extract the image features using various filters like Gaussian filters etc. at the beginning, and the CNN models also extract extensive features using large number of convolutional filters. Secondly, the transform domain fusion algorithms usually fuse the features through the weighted average strategy, and the CNN models also utilize the weighted average strategy to generate the target image. On comparing the transform domain based image fusion and CNN based image fusion models, the CNN models have three advantages: 1) the parameters of convolutional filters can be learned to fit the image fusion task; 2) the number of convolutional filters is usually much greater than that of the filters in the convolutional transform domain algorithms, and thus the convolutional filters could extract more informative image features; 3) the parameters of the CNN models can be cooperatively optimized through training them in the end-to-end manner. Liu et al. [19] calculates the three commonly

used focus measures (feature extraction) and them feeds them to a three-layer network (input-hidden-output), so the network just acts as a classifier for the fusion rule design. As a result of which, the source images must be fused patch by patch. Here, the CNN model is simultaneously used for activity level measure (feature extraction) and fusion rule design (classification).

II. CNN for Image Fusion

CNN is a deep learning, trainable, multi-stage feed forward artificial neural network, which attempts to learn hierarchical feature representation mechanism, and each level of CNN contains a certain number of feature maps corresponding to a level of abstraction for features. The operations such as non-linear activation, linear convolution and spatial pooling applied to coefficients are used to connect the feature maps at different stages. In past few years, CNN has been successfully introduced into various fields in computer vision from high-level tasks to low-level tasks, such as face detection [19], semantic segmentation [20], face recognition [21], super-resolution [22], patchy comparison [23] etc. These CNN based methods usually overtake the convolutional methods in their respective fields, possessing to the fast development of modern powerful GPUs, the great progress on effective training techniques, and the easy access to a large amount of image data. This work also benefits from these factors.

a) Overview

The basic working diagram of CNN based multi-focus image fusion method is shown in figure 1, in this the main consideration is the situation where there are only two pre-registered source images. To deal with more than two multi-focus images, one can fuse them one by one in series. From the figure, it is understandable that this method consists of four steps: focus detection, initial segmentation, consistency verification and fusion. In the first step, the two source images are fed to a pre-trained CNN model to output a score map, which contains the focus information of source images. Predominantly, each coefficient in the score map indicates the focus property of a pair of corresponding patches from two source images. Then, a focus map with the same size of source images is obtained from the score map by averaging the overlapping patches. In the second step, the focus map is segmented into a binary map with a threshold of 0.5. In the third step, refine the binary segmented map with two popular consistency verification strategies, namely, small region removal and guided image filtering [24], to generate the final decision map. In the last step, the fused image is obtained with the final decision map using the pixel-wise weighted-average strategy.

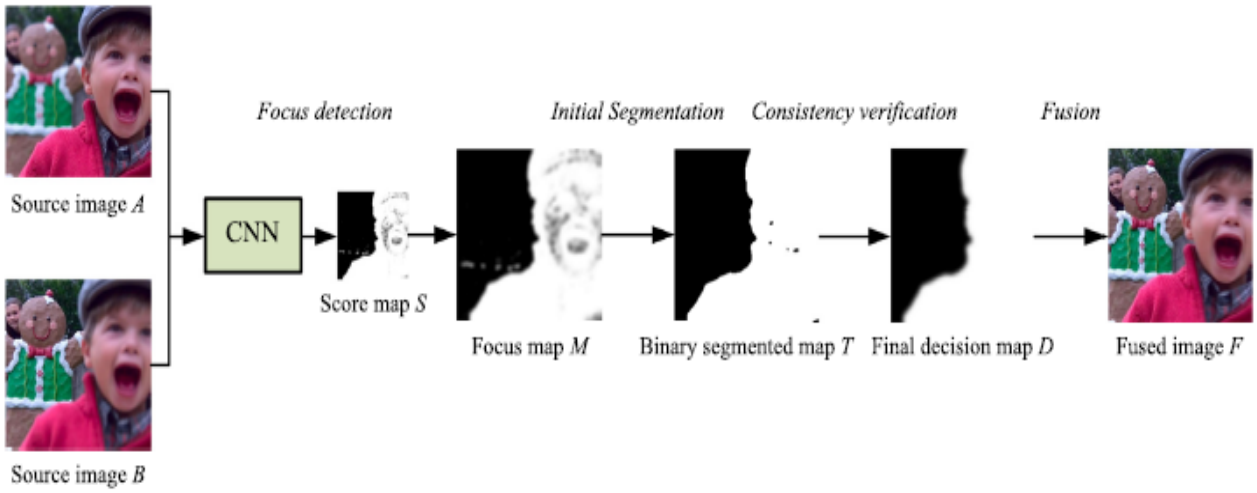


Figure 1: Diagram of the CNN based Multi-focus Image Fusion Method [8]

b) Design

In this work, multi-focus image fusion is viewed as a two-class classification problem. For a pair of image patches {*p1, p 2*} of the same scene, the goal is to learn a CNN whose output is a scalar ranging from 0 to 1. Specifically, the output value should be close to 1 when *p1* is focused while *p2* is defocused, and the value should be close to 0 when *p2* is defocused while *p1* is focused. In other words, the output value indicates the focus property of the patch pair. To this end, here employ a large number of patch pairs as training examples. Each training example is a patch pair of the same scene. One training example {*p* 1, *p* 2} is defined as a positive example when *p* 1 is clearer than *p* 2, and its label is set to 1. On the contrary, the example is defined as a negative example when *p* 2 is clearer than *p* 1 and the label is set to 0. In practical usage, the source images have arbitrary spatial size. One possible way is to apply sliding-window technique to divide the images into overlapping patches, and then input each pair of patches into the network to obtain a score. However, considering that there are a large number of repeated convolutional calculations since the patches are greatly overlapped, this patch-based manner is very time consuming. Another approach is to input the source images into the network as a whole without dividing them into patches, aiming to directly generate a dense prediction map. Since the fully-connected layers have fixed dimensions on input and output data, to make it possible, the fully-connected layers should be firstly converted into convolutional layers by reshaping parameters. After the conversion, the network only consists of convolutional and max- pooling layers, so it can process source images of arbitrary size as a whole to generate dense predictions [25]. As a result, the output of the network now is a score map, and each coefficient within it indicates the focus property of a pair of patches in source images. The patch size equals to the size of training examples. When the kernel stride of each convolutional layer is one pixel, the stride of adjacent patches in source images will be just determined by the number of max-pooling layers in the network. To be more specific, the stride is 2 *k* when there are totally *k* max-pooling layers and each with a kernel stride of two pixels [20].

There are three types of CNN models are presented for patch similarity comparison: *siamese, pseudo-siamese* and *2-channel*. The siamese network and pseudo-siamese network both have two branches with the same architectures, and each branch takes one image patch as input. The difference between these two networks is the two branches in the former one share the same weights while in the latter one do not. Thus, the pseudo-siamese network is more flexible than the siamese one. In the 2-channel network, the two patches are concatenated as a 2-channel image to be fed to the network. The 2-channel network just has one trunk without branches. Clearly,

for any solution of a siamese or pseudo-siamese network, it can be reshaped to the 2-channel manner, so the 2-channel network provides further more flexibility [23]. All the above three types of networks can be adopted in the proposed CNN-based image fusion method. In this work, the siamese one is used as here the CNN model mainly for the following two considerations. First, the siamese network is more natural to be explained in image fusion tasks. The two branches with same weights demonstrate that the approach of feature extraction or activity level measure is exactly the same for two source images, which is a generally recognized manner in most image fusion methods. Second, a siamese network is usually easier to be trained than the other two types of networks. As mentioned above, the siamese network can be viewed as a special case of the pseudo-siamese one and 2-channel one, so its solution space is much smaller than those of the other two types, leading to an easier convergence. Another important issue in network design is the selection of input patch size. When the patch size is set to 32 ×32, the classification accuracy of the network is usually higher since more image contents are used. However, there are several defects which cannot be ignored using this setting. As is well known, the max- pooling layers have important significance to the performance of a convolutional network. When the patch size is 32 ×32, the number of max-pooling layers is not easy to determine. More specifically, when there are two or even more max-pooling layers in a branch, which means that the stride of patches is at least four pixels, the fusion results tend to suffer from block artifacts. On the other hand, when there is only one max-pooling layer in a branch, the CNN model size is usually very large since the number of weights in fully-connected layers significantly increases. Further- more, for multi-focus image fusion, the setting of 32 ×32 is often not very accurate because a 32 ×32 patch is more likely to contain both focused and defocused regions, which will lead to undesirable results around the boundary regions in the fused image. When the patch size is set to 8 ×8, the patches used to train a CNN model is too small that the classification accuracy cannot be guaranteed. Based on the above considerations as well as experimental tests, we set the patch size to 16 ×16 in this study.

Fig. 2 shows the CNN model used in the proposed fusion algorithm. It can be seen that each branch in the network has three convolutional layers and one max-pooling layer. The kernel size and stride of each convolutional layer are set to 3 ×3 and 1, respectively. The kernel size and stride of the max-pooling layer are set to 2 ×2 and 2, respectively. The 256 feature maps obtained by each branch are concatenated and then fully-connected with a 256-dimensional feature vector. The output of the network is a 2-dimensional vector that is fully-connected with the 256- dimensional vector. Actually, the 2-dimensional vector is fed to a 2-way soft max layer (not shown in Fig. 2) which produces a probability distribution over two classes. In the test/fusion process, after converting the two fully-connected layers into convolutional ones, the network can be fed with two source images of arbitrary size as a whole to generate a dense score map [39,43,45]. When the source images are of size $H \times W$, the size of the output score map is $(\lceil H/2^1 \rceil - 8 + 1) \times (\lceil W/2^1 \rceil - 8 + 1)$, where $\lceil\ \rceil$ denotes the ceiling operation. Each coefficient in the score map keeps the output score of a pair of source image patches of size 16 ×16 going forward through the network. In addition, the stride of the adjacent patches in source images is two pixels because there is one max-pooling layer in each branch of the network.

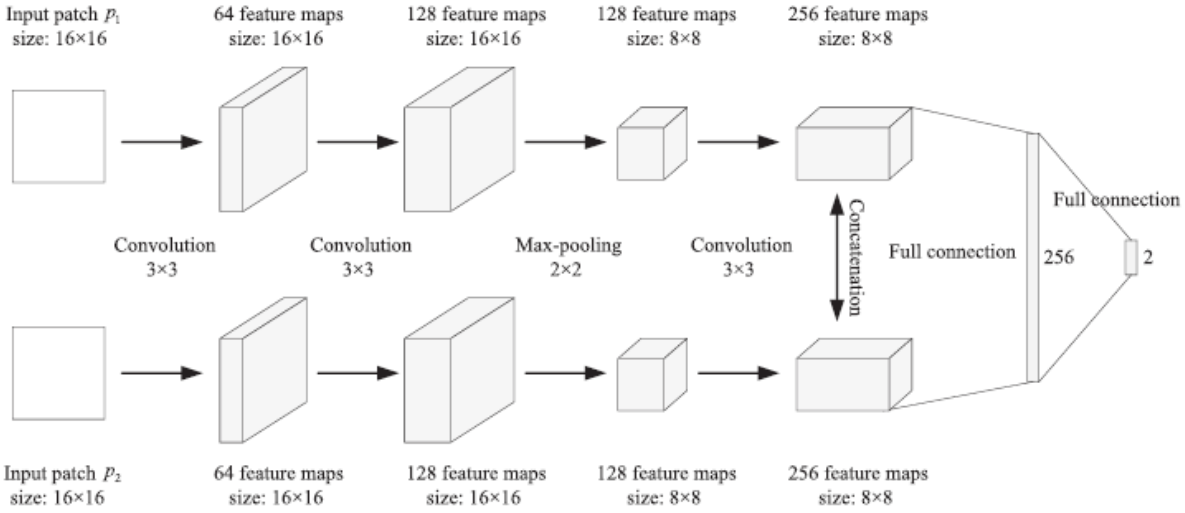


Figure 2: CNN model for fusion process (notice that the spatial size marked in the figure just indicates the training process)

c) Training

The training examples are generated from the images in image set, which contains high quality natural images. For each image (converted into grayscale space at first), five blurred versions with different blurring level are obtained using Gaussian filtering. Specifically, a Gaussian filter with a standard deviation of 2 and cut off to 7 ×7 is adopted here. The first blurred image is obtained from the original clear image with the Gaussian filter. The second blurred image is obtained from the first blurred image with the filter, and so on. Then, for each blurred image and the original image, 20 pairs of patches of size 16 ×16 are randomly sampled (the patch sampled from the original image must has a variance larger than a threshold, e.g., 25). Let $p\ c$ and $p\ b$ denote a pair of clear and blurred patches, respectively. It is defined as a positive example (label is set to 1) when $p\ 1 = p\ c$ and $p\ 2 = p\ b$, where $p\ 1$ and $p\ 2$ are the input of the first and second branch respectively. On the contrary, it is defined as a negative example (label is set to 0) when $p\ 1 = p\ b$ and $p\ 2 = p\ c$. Thus, the training set finally consists of few positive examples and few negative examples. As with CNN-based classification tasks [21-26], the softmax loss function (multinomial logistic loss of the output after applying softmax) is used as the objective of our network. The stochastic gradient descent (SGD) is applied to minimize the loss function. The weights are updated with the following rule

$$v_{i+1} = 0.9 \cdot v_i - 0.0005 \cdot \alpha \cdot w_i - \alpha \cdot \partial L / \partial w_i, \qquad w_i + 1 = w_i + v_{i+1},$$

where v is the momentum variable, I is the iteration index, α is the learning rate, L is the loss function, and $\partial L/\partial w_i$ is the derivative of the loss with respect to the weights at w_i. CNN model is trained using the popular deep learning framework Caffe [28]. The weights of each convolutional layer are initialized with the Xavier algorithm [29], which adaptively determines the scale of initialization according to the number of input and output coefficients. The biases in each layer are initialized as 0. The leaning rate is equal for all layers and initially set to 0.0001. The learning rate manually drop by a factor of 10 when the loss reaches a stable state. The trained network is finally obtained after about 10 epochs through the 2 million training examples. The learning rate is dropped one

time throughout the training process. One may notice that the training examples could be sampled from real multi-focus image dataset rather than just artificially created via Gaussian filtering. Of course, this idea is good and feasible. Actually, we experimentally verify this idea by building another training set in which half of the examples originate from a real multi-focus image set while the other half are still obtained by the Gaussian filtering based approach. we also construct a validation set which contains 10,000 patch pairs from some other multi-focus images for verification. The result shows that the classification accuracies using the above two training set with same training process are approximately the same, both around 99.5% (99.49% for the pure Gaussian filtering based set while 99.52% for the mixed set). Moreover, from the viewpoint of final image fusion results, the difference between these two approaches is even smaller that can be neglected. This test indicates that the classifier trained by the Gaussian filtering based examples can tackle the defocus blur very well. An explanation about it is that in our opinion, as the Gaussian blur is conducted on five different standard deviations, the trained classifier could handle most blur situations, which is not limited to the situations of five discrete standard deviations in the training set, but greatly expanded to a lot of combinations (may be linear or nonlinear) of them. Therefore, there is a very large possibility to cover the situations of defocus blur in multi- focus photography. To verify it, here applied a new training set which consists of Gaussian filtered examples using only three different standard deviations, and the corresponding classification accuracy on the validation set has a remarkable decrease to 96.7%. Furthermore, there is one benefit when using this artificially created training set. That is, I can naturally extend the learned CNN model to other-type image fusion issues, such as multi-modal image fusion and multi-exposure image fusion. Otherwise, when the training set contains examples sampled from multi-focus images, this extension seems to be not reasonable. Thus, the model learned from artificially created examples tends to have a stronger ability of generalization. To have some insights into the learned CNN model, here provided some representative output feature maps of each convolutional layer. The example images shown in Fig. 1 are used as the inputs. For each convolutional layer, two pairs of corresponding feature maps (the indices of two branches are the same) are shown in Fig. 3. The values of each map are normalized to the range of [0, 1]. For the first convolutional layer, some feature maps captures high- frequency information as shown in the left column while some others are similar to the input images as shown in the right column. This indicates the spatial details cannot be fully characterized by the first layer. The feature maps of the second convolutional layers mainly concentrate on the extraction of spatial details covering various gradient orientations. As shown in Fig. 3, the left and right columns mainly capture horizontal and vertical gradient information, respectively. This gradient information is integrated by the third convolutional layer, as its output feature maps successfully characterize the focus information of different source images. Accordingly, with the following two fully-connected layers, an accurate score map could be finally obtained.

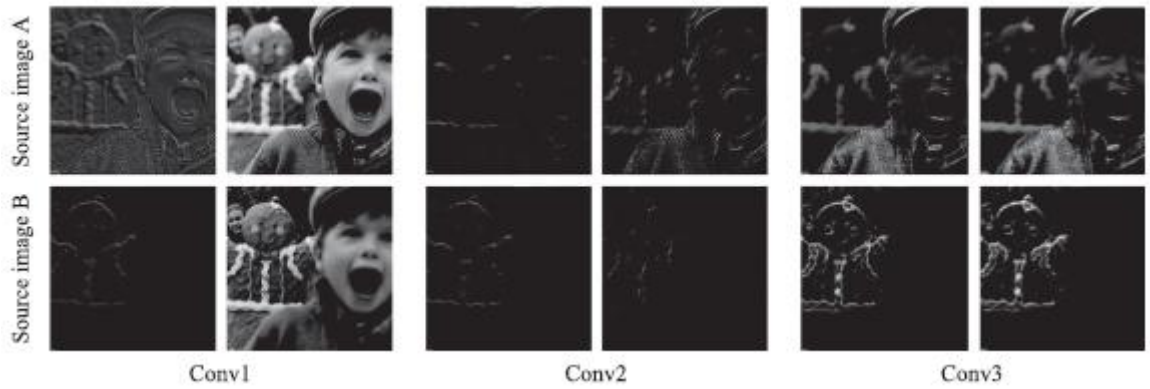


Figure 3: Representative output feature maps of each convolutional layer

d) Fusion Scheme

Focus Detection: Let A and B denote the two source images. In the proposed fusion algorithm, the source images are converted to grayscale space if they are colour images. Let $\hat{A}$ and $\hat{B}$ denote the grayscale version of A and B (keep $\hat{A} = A$ and $\hat{B} = B$ when the source images are originally in grayscale space), respectively. A score map S is obtained by feeding $\hat{A}$ and $\hat{B}$ to the trained CNN model. The value of each coefficient in S ranges from 0 to 1, which suggests the focus property of a pair of patches of size 16 ×16 in source images. The closer the value is to 1 or 0, the more focused the patch from source image $\hat{A}$ or $\hat{B}$ is. For two neighbouring coefficients in S, their corresponding patches in each source image are overlapped with a stride of two pixels. To generate a focus map (denoted as M) with the same size of source images, we assign the value of each coefficient in S to all the pixels within its corresponding patch in M and average the overlapping pixels. Fig. 4(a) shows the obtained focus map of the example illustrated in Fig. 1. It can be seen that the focus information is accurately detected. Intuitively, the values of the regions with abundant details seems to be close to 1 (white) or 0 (black), while the plain regions tend to own values close to 0.5 (gray).

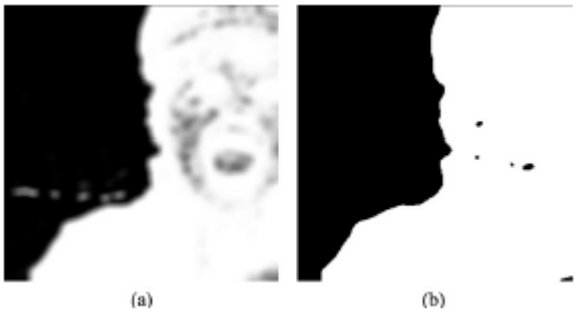


Figure 4: Initial Segmentation (a) Focus Map (b) Binary Segmentation Map

Initial Segmentation: To preserve useful information as much as possible, the focus map M needs to be further processed. In this method, as with most spatial domain multi-focus image fusion methods [25,27,30-33], here we also adopt the popular "choose-max" strategy to process M. Accordingly, a fixed threshold of 0.5 is applied to segment M is into a binary map T,

which is in accord with the classification principle of the learned CNN model. The obtained binary map is shown in Fig. 4(b) (please notice the optical illusion in the focus map shown in Fig. 4(a), namely, the gray regions seems to be darker than its real intensity in a white background while brighter than its real intensity in a black back- ground). It can be seen that almost all the gray pixels in the focus map are correctly classified, which demonstrates that the learned CNN model can obtain precise performance even for the plain regions in source images.

Consistency Verifications: It can be seen from Fig. 4(b) that the binary segmented map is likely to contain some misclassified pixels, which can be easily removed using the small region removal strategy. Specifically, a region which is smaller than an area threshold is reversed in the binary map. One may notice that the source images sometimes happen to contain very small holes. When this rare situation occurs, users can manually adjust the threshold even to zero, which means the region removal strategy is not applied. In this paper, the area threshold is universally set to $0.01 \times H \times W$, where H and W, are the height and width of each source image, respectively.

Figure 5: (a) Initial Decision Map (b) Initial Fused Image (c) Final Decision Map (d) Fused Image

Fig. 5(a) shows the obtained initial decision map after applying this strategy. Fig. 5(b) shows the fused image using the initial decision map with the weighted-average rule. It can be seen that there are some undesirable artifacts around the boundaries between focused and defocused regions. Similar to [27], we also take advantage of the guided filter to improve the quality of initial decision map. Guided filter is a very efficient edge-preserving filter, which can transfer the structural information of a guidance image into the filtering result of the input image. The initial fused image is employed as the guidance image to guide the filtering of initial decision map. There are two free parameters in the guided filtering algorithm: the local window radius r and the regularization parameter ε. In this work, we experimentally set r to 8 and ε to 0.1. Fig. 5(c) shows the filtering result of the initial decision map given in Fig. 5(b).

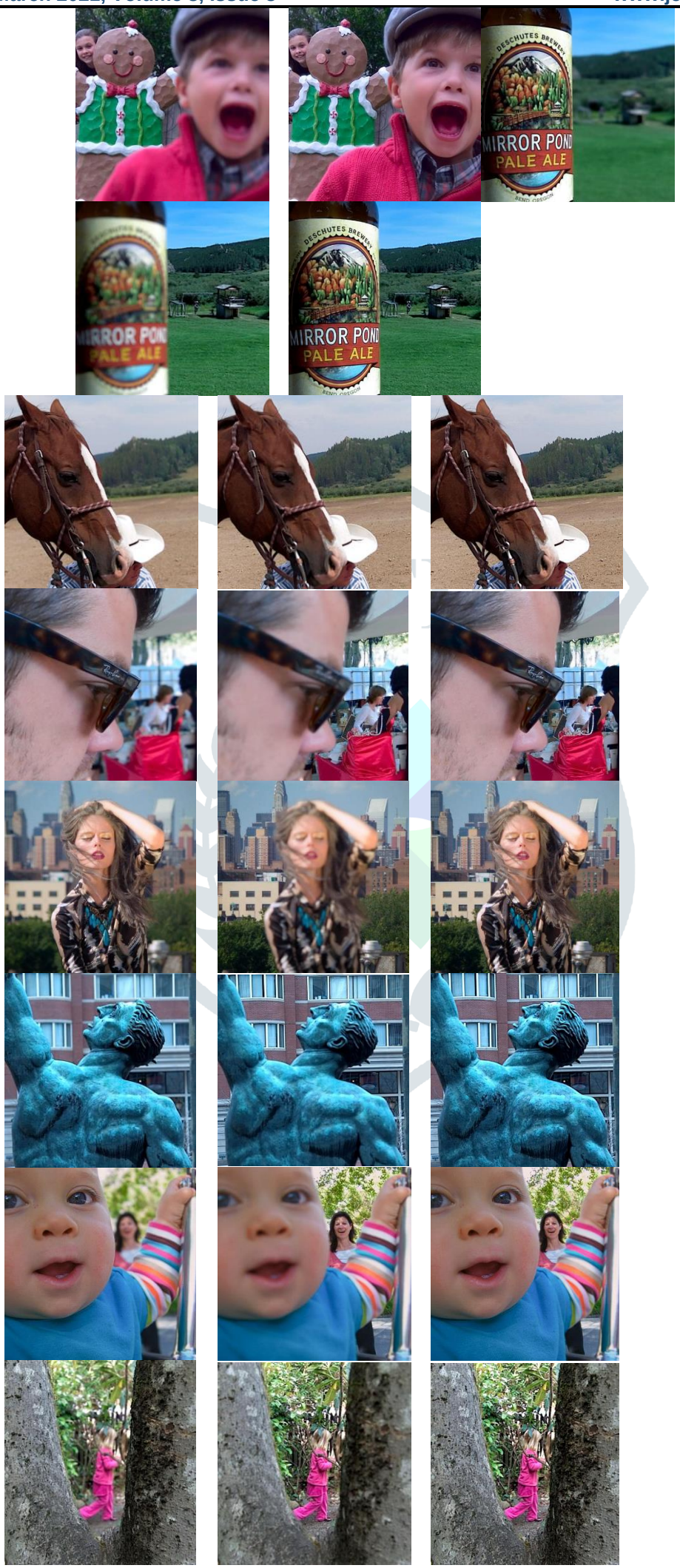
Fusion: Finally, with the obtained decision map D, at last calculate the fused image F with the following pixel-wise weighted-average rule

$$F(x, y) = D(x, y)\, A(x, y) + (1 - D(x, y))\, B(x, y)$$

The fused image of the given example is shown in Fig. 5(d).

III. Result

Since the model has been only trained on the multi-focus image dataset, thus at first investigate the performance of the proposed model on fusing multi-focus images. Additionally, we also want to test the effectiveness of the perceptual loss and fusion rules on this dataset. In order to achieve the above two purposes, the algorithm is evaluated on the multi-focus image dataset.



(a) (b) (c)

Figure 6: (a) Input Image A (b) Input Image B (c) Fused Image

Fig. 6 displays an example of the Lytro dataset, which comprises the fusion results of multifocus image fusion algorithm based on deep learning convolutional neural network method. From the result image of CNN algorithm, it can be clearly seen that the fusion quality is better, because it uses many processing operations. The main novelty of our method is learning a CNN model to achieve a direct mapping between source images and the focus map. Based on this idea, the activity level measurement and fusion rule can be jointly generated by learning the CNN model, which can overcome the difficulty faced by the existing fusion methods. CNN mapping process starting from source images to the focus map, which is the common core task of various image fusion issues as this mapping process simultaneously involves activity level measurement and comparison (namely, fusion rule). The subsequent techniques applied to the focus map could be selected or designed according to the characteristics of a specific fusion task. This is a reasonable way to study this topic from our perspective as we believe conventional techniques in related fields are still of high value and should not be discarded. In this work, we just employ some popular techniques for multi-focus image fusion issues, so further studies following this route could be performed in the future.

IV. Conclusion

In this paper, a general method of multifocus image fusion based on convolutional neural network is proposed. The main originality of this method is learning a CNN model to achieve a direct mapping between source images and the focus map. Based on this idea, the activity level measurement and fusion rule can be jointly generated by learning the CNN model, which can overcome the difficulty faced by the existing fusion methods. Finally, this model is designed as a general image fusion structure, thus its performance might be limited for fusing a specific type of images. Therefore, one practical way, to improve performance of the CNN based image fusion models, is to design the architecture according to the specific characteristics of the target image dataset.

In future work, we should focus on other image fusion fields, such as medical image fusion. We should design the end-to-end network structure based on the imaging characteristics of medical images. Medical image data sets are not like multi-focus images that can be acquired using natural images, so we have to design a special medical image-based data set for medical image fusion.

References

[1] Hui Li, Xiao-Jun Wu, Josef Kittler, "Infrared and Visible Image Fusion using a Deep Learning Framework," arXiv:1804.06992v4[cs.CV], Dec 2018.

[2] B. Rajalingam, R. Priya, "Multimodal Medical Image Fusion based on Deep Learning Neural Network for Clinical Treatment Analysis," International Journal of ChemTech Research, 2018,11(06): 160-176.

[3] T. Wan, C. Zhu, and Z. Qin, "Multifocus image fusion based on robust principal component analysis," Pattern Recognition Letters, vol. 34, no. 9, pp. 1001–1008, 2013.

[4] Z. Zhong, T. Shen, Y. Yang, Z. Lin, and C. Zhang, "Joint subbands learning with clique structures for wavelet domain super-resolution," *Advances in Neural Information Processing Systems*, 2018.

[5] Jinjiang Li, Genji Yuan, Hui Fan, "Multifocus Image Fusion Using Wavelet Domain based Deep CNN," Hidawi Computational Intelligence and Neuroscience Volume 2019, Article ID 4179397.

[6] Zhe Guo , Xiang Li , Heng Huang, Ning Guo, and Quanzheng Li, "Deep Learning based Image Segmentation on Multimodal Medical Imaging," IEEE Transactions on Radiation and Plasma Medical Sciences, Vol. 3, No. 2, March 2019.

[7] Jingchun Piao, Yunfan Chen, Hyunchul Shin, "A New Deep Learning Based Multi-Spectral Image Fusion Method," Entropy 2019, 21, 570.

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

[9] K. Xu, Z. Qin, G. Wang, H. Zhang, K. Huang, S. Ye, "Multi-focus Image Fusion using Fully Convolutional two-stream Network for visual sensrs," KSII Transactions on Internet and Information System, vol. 12, no. 5, pp. 2253-2272, 2018.

[10] L. Zhao, H. Bai, A. Wang, and Y. Zhao, "Multiple Description convolutional neural networks for image compression," IEEE Transactions on Circuits and Systems for Video Technology, 2018

[11] C. Mingrui, Y. Junyi, C. Guanghui, "Multi-focus Image Fusion Algorithm using LP Transformation and PCNN," in Proceedings of 2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS), pp.237-241, September 2015.

[12] B. Yang, S. Li, Multifocus Image Fusion and Restoration with Sparse Representation, IEEE Trans. Instrumen Meas. 59 (4) (2010) 884-892.

[13] Y.Liu, X. Chen, Z. Wang, Z.J. Wang, R.K. Ward, X. Wang, "Deep Learning for Pixel Level Image Fusion: Recent and Advances and Future Prospects," Inf. Fusion 42 (2018) 158-173.

[14] Y. Liu, X. Chen, Z. Wang, H. Peng, "Multi-focus Image Fusion with a Deep Convolutional Neural Network," Inf. Fusion 36 (2017) 191-207.

[15] H. Tang, B. Xiao, W. Li, G. Wang, "Pixel Convolutional Neural Network for Multi-Focus Image Fusion," Inf. Science (2017).

[16] H. Song, Q. Liu, G. Wang, R. Hang, B. Huang, "Spatiotemporal Satellite Image Fusion using Deep Convolutional Neural Network," IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 11 (3) (2018) 821-829.

[17] Z. Zhoni, B. Wang, S.Li, M. Dong, "Perceptual Fusion of Infrared and Visible Images through a Hybrid Multi-Scale Decomposition with Gaussian and Bilateral Filters," Inf. Fusion 30 (2016) 15-26.

[18] S. Li, J. Kwok, Y. Wang, "Multifocus Image Fusion using Artificial Neural Networks," Pattern Recognition Letter 23 (8) (2002) 985-997

[19] S. Farfade, M. Saberian, L. li, "Multiview Face Detection using Convolutional Neural Networks," Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, (2015) 643-650.

[20] J. Long, E. Shelhamer, T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (2015) 3431-3440.

[21] Y. Sun, X. Wang, X. Tang, "Deep Learning Face Representation from Predicting" Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (2014) 1891-1898.

[22] C. Dong, C. Loy, K. He, X. Tang, "Image Super-Resolution using Deep Convolutional Netwroks," IEEE Trans. Pattern Analysis Mach. Intell. 38 (2) (2016) 295-307.

[23] S. Zagoruvko, N. Komodakis, "Learning to compare Image Patches via Convolutional Neural Network," Proceesings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, 4353-4361.
[24] K. He, J. Sun, X. Tang, "Guided image filtering," IEEE Trans. Pattern Anal. Mach. Intell. 35 (6) (2013) 1397–1409.
[25] S. Li, X. Kang, J. Hu, "Image fusion with guided filtering," IEEE Trans. Image Process. 22 (7) (2013) 2864–2875.
[26] A. Krizhevsky, I. Sutskever, G. Hinton, "Imagenet classification with deep convolutional neural networks," in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105
[27] M. Nejati, S. Samavi, S. Shirani, "Multi-focus image fusion using dictionary-based sparse representation," Inf. Fusion 25 (1) (2015) 72–84.
[28] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadar-rama, T. Darrell, "Caffe: convolutional architecture for fast feature embedding," in: Proceedings of the ACM International Conference on Multimedia, 2014, pp. 675–678.
[29] X. Glorot, Y. Bengio, "Understanding the difficulty of training deep feed forward neural networks," in: International Conference on Artificial Intelligence and Statistics, 2010.
[30] Y. Liu, S. Liu, Z. Wang, "Multi-focus image fusion with dense sift," Inf. Fusion 23 (1) (2015) 139–155.
[31] S. Li, B. Yang, "Multifocus image fusion using region segmentation and spatial frequency," Image Vis. Computation 26 (7) (2008) 971–979.
[32] S. Li, X. Kang, J. Hu, B. Y, "Image matting for fusion of multi-focus images in dynamic scenes," Inf. Fusion 14 (2) (2013) 147–162
[33] Z. Zhou, S. Li, B. Wang, "Multi-scale weighted gradient-based fusion for multi–focus images," Inf. Fusion 20 (1) (2014) 60–72.