Super-Resolution multi-focus image fusion based on convolutional neural network

Wang Ying¹,a, Zeng Shangyou¹,b *, Wang Xinjiao¹,c, Wang Jinjin¹,d

¹ School of Electronic Engineering, Guangxi Normal University, Guilin, Guangxi, China, 541004

aemail: wasdfgy@163.com, cemail: 1184902443@qq.com, demail: 1516844863@qq.com

Corresponding author’s email: bemail: zsy@mailbox.gxnu.edu.cn

Abstract: Multi-focus image fusion combines the focus parts of multiple images in the same scene to generate a fully focused image. Aiming at the problem of multi-focus image fusion, this paper optimizes the network design based on the existed image fusion network. The optimized network uses two groups of dilated convolutions with different numbers and sizes to extract shallow features, and then integrates the two groups of features through 1 × 1 convolution. Extracting deep information from feature map by dense block and the dense block are composed of four layers of dense connection convolution. In addition, the fused image is processed by sub-pixel convolution to obtain a super-resolution fused image, and a perceptual loss function is introduced to optimize the network. The experimental results show that the optimized network is superior to the existing fusion methods in objective and subjective evaluation. It has a certain application value in the field of image fusion.

1. Introduction

Due to the limitation of optical imaging depth of field, it is a challenge to capture a clear image of all objects in a scene. The main reason is that only objects at a certain distance from the camera can focus and be clearly captured, while objects at other distances in front of or behind the focal plane will defocus and blur. Multi-focus image fusion refers to the fusion of images with different focal lengths to generate a single composite image. The fused image provides better visual perception and more information, which is suitable for computer vision and image processing tasks, such as target detection, image segmentation etc.

Various image fusion methods have been proposed, which can be roughly divided into transform domain method and spatial domain method. The spatial domain fusion method based on image fusion uses pixel level gradient information or image blocks for image fusion, but this process will introduce block effect to affect the fusion effect. Common spatial domain methods such as pixel-level fusion methods is based on dense scale invariant feature transform (DSIFT)[1], the guide filtering fusion (GFF)[2] and uniform similarity[3]. Different from the spatial domain fusion method, the method based on transform domain, such as using different multi-scale transform (MST), decomposes the source image into sub images with different resolution layers or different frequency bands. It can effectively avoid the blocking effect and improve the fusion efficiency. However, the transform domain method needs down sampling operation, which will lead to more or less artifacts in the fused image. Some methods based on transform domain include dual tree complex wavelet transform (DTCWT)[4], sparse representation (SR)[5].
Convolutional neural network (CNN), as an important branch of deep learning, can learn to use hierarchical structure to represent the input image. Therefore, CNN has stronger feature extraction ability than traditional methods and is naturally applied in the field of image fusion. Liu et al. [6] proposed an image fusion method based on deep convolution neural network for the first time. They take multi-focus image fusion as a classification task (focus or blur), and use CNN to predict the focus image. The architecture of image fusion model based on CNN still has a lot of room to improve their performance and generalization.

2. Densefuse network
Densefuse [7] is a fusion network for processing infrared and visible images. The network is divided into three parts: feature extraction module, feature fusion module, feature reconstruction module. The feature extraction part uses 3 × 3 convolution to extract shallow features, and the three dense connected convolutions to extract deep features. The dense block all use 3 × 3 convolution. The feature fusion part uses additive strategy and L1 norm [8] strategy, and the fusion strategy structure based on L1 norm is shown in Figure 1. In the feature reconstruction part, four 3 × 3 convolutions are used for image reconstruction.

Strategy of addition:

\[ f^m(x, y) = \sum_{i=1}^{k} \omega^m_i(x, y) \]  

(1)

Where \( (x, y) \) denotes the corresponding position in the feature graph and fusion feature graph of feature extraction, \( \omega^m_i(x, y) \) represents the feature graph of feature extraction, \( f^m \) represents the fusion feature graph.

Strategy of L1 norm:

\[ f^m(x, y) = \sum_{i=1}^{k} \omega_i(x, y) \times \omega^m_i(x, y) \]  

\[ \omega_i(x, y) = \frac{C_i(x, y)}{\sum_{n=1}^{k} C_n(x, y)} \]  

(3)

Where \( \omega^m \) is the characteristic graph, \( C_i \) is the initial fusion feature map, \( C_i \) is calculated by L1 norm and block based averaging operator, and \( f^m \) represents the final fusion feature map.

3. Optimize the network
This paper optimizes Densefuse network to deal with multi-focus image fusion. Since additive fusion is a very rough fusion strategy for salient features, the fusion strategy in this paper only adopts L1 norm fusion strategy, and the optimized network structure is shown in Figure 2. In this paper, feature extraction, feature reconstruction and loss function are optimized.
3.1 Feature extraction

Unlike Densefuse, which uses a single size convolution kernel for shallow feature extraction, we use two groups of $3 \times 3$ dilated convolution with different numbers and expansion coefficients to extract features, and the expansion coefficients of dilated convolution are 1 and 2 respectively. The extracted features are cascaded and integrated by $1 \times 1$ convolution. Dilated convolution can expand the receptive field without increasing the size of convolution kernel, and the combination of different size dilated convolution can extract more detailed features. In order to further enhance the information interaction between layers, we deepen a layer of dense connection convolution. Standard convolution operation and dilated convolution operation can be uniformly formulated as

$$O(j, k) = \sum \sum W(m, n) \cdot I(j + m \cdot d, k + n \cdot d)$$

(4)

where $d$ stands for dilation rate. When $d$ is set to 1, the formula represents a standard convolution operation.

3.2 Feature reconstruction

Generally speaking, the higher the resolution of the image, the more details it contains and the more information it contains. We regard fusion feature image as low resolution image, and feature reconstruction can be regarded as the process of generating super-resolution image from low resolution image. Therefore, our feature reconstruction uses sub-pixel convolution [9] to obtain super-resolution image.

We continue to use the three convolution layers of Densefuse network to reconstruct the image. After three convolutions of fused features, the channel number of feature map is $R^2$ and the size is $H \times W$. The last layer convolution uses sub-pixel convolution, which makes the $R^2$ channels of each pixel of the feature image arranged and combined into an $R \times R$ region. At this time, each $R \times R$ region corresponds to a sub block of $H \times W \times R^2$ size in the super-resolution image, so the feature image of $H \times W \times R^2$ is recombined into a super-resolution image of $RH \times RW \times 1$. The reconstruction module convolution is shown in Figure 3.
3.3 Loss function

The same images are offset by one pixel from each other, and although they are similar in perception, they are quite different in terms of pixel loss. Perceptual loss\cite{10} measures image similarity more robustly than pixel loss. In the optimized network, the perceptual loss function is added to the original pixel loss and structural similarity loss to form a new loss function. The formula is as follows:

$$L = \alpha L_p + \beta L_{SSIM} + \gamma T_{LOSS}$$  \hspace{1cm} (5)

Where $L_p$ is the pixel loss, $L_{SSIM}$ is the pixel loss measuring the similarity of image structure, $T_{LOSS}$ represents perceived loss. $\alpha$, $\beta$ and $\gamma$ represent the weight coefficients of pixel loss, image structure similarity and perception loss respectively.

The pixel loss is calculated as follows:

$$L_p = ||O - I||_2$$  \hspace{1cm} (6)

Where $O$ and $I$ represent the output and ground truth images respectively. It is the Euclidean distance between output $O$ and input $I$.

$$L_{SSIM} = 1 - SSIM(O, I)$$  \hspace{1cm} (7)

Where $SSIM(\cdot, \cdot)$ \cite{11} represents the structural similarity operation, which represents the structural similarity of two images.

This paper uses the features of the last convolution layer of ResNet101, which is pretrained on the Imagenet dataset to construct the perceptual loss. Specifically, the proposed perceptual loss is expressed as the mean square error of the feature mapping between the predicted fusion image extracted from the last convolution layer of resnet101 and the ground truth fusion image.

$$T_{LOSS} = \frac{1}{c_{HF}W_F} \sum_{i,j} (f_i(x,y) - f_0(x,y))^2$$  \hspace{1cm} (8)

Where $f_i$ and $f_0$ is the feature map of prediction fusion image and ground truth fusion image, $i$ is the channel index of feature map, $C_F$, $H_F$, $W_F$ is the channel number, height and width of the feature map.

4. Experiment

4.1 Experimental setup

Experimental environment: this experiment uses Python as the underlying framework of deep learning, the computer memory is 32GB RAM, Intel i7-6700k 4 core 8 thread CPU and NVIDIA-GTX 1080ti GPU, and the operating system is windows 10 64 bit.

In the training phase, we use MS-COCO dataset as training data. MS-COCO dataset is a popular open source image dataset, which is widely used for classification and segmentation tasks. Among these training images, about 80000 images are used as input images, and 1000 images are used to verify the reconstruction ability of each iteration. All images are adjusted to $256 \times 256$ and converted to gray images. When the learning rate is set to $1 \times 10^{-5}$, the learning rates of all layers are equal. The whole network uses Adam optimizer to minimize the loss function. The batch size and iterations respectively were 2 and 4.

In the following experiments, the optimized network will be compared with SR, GFF, DSIFT, Densefuse and ECNN\cite{12} based on deep learning. Firstly, the visual effect of the fused image relative to the source image is qualitatively evaluated. Specifically, multi-focus image fusion should integrate as many clear features of each source image into the fusion image as possible, so that the human eye can easily and accurately obtain comprehensive information from the fusion image. Secondly, for a more comprehensive and objective comparison, we use the following four quantitative indicators to quantitatively evaluate the performance of the fusion method from different aspects. The four quantitative indicators are: structural similarity ($SSIM_a$), Edge information quality ($Q_{AB/F}$) \cite{13}, Tsallis entropy ($Q_{TE}$) \cite{14} and mutual information ($Q_{MI}$)\cite{15}. These four metrics can effectively reflect the ability of the model to fuse visual information, structural information and image details. In our experiment, the formula of $SSIM_a$ is as follows:
\[
SSIM_a(F) = \left( SSIM(F, I_1) + SSIM(F, I_2) \right) \times 0.5
\]  

(9)

Where \( SSIM(\cdot) \) denotes structural similarity operation, \( F \) denotes fusion image, \( I_1 \) and \( I_2 \) represents the source image. The fusion performance improves with the increasing numerical index of all these seven metrics.

4.2 Evaluation of fusion methods

In this section, Figure 4 shows the fusion result of an example in the "Lytro" dataset, which includes transform domain method SR, GFF, spatial domain method DSIFT, Densefuse, ECNN and the optimization model based on deep learning.

As can be seen from Figure 4, although SR algorithm has good human visual effect, there are redundant fuzzy artifacts near the boundary of the wire fence. The fused image of GFF algorithm has slight blocking effect in the wire fence. The wire fence of DSIFT algorithm has obvious block effect. The Densefuse based method produces redundant blur, and the fused image shows severe color distortion, which is obviously not in line with human visual habits. ECNN fusion image can clearly observe the block effect and jagged edges in the wire fence. The fusion image of the optimization network in this paper well fuses the clear features of the two source images, and shows good human visual effect.

| Metric      | SR    | GFF   | DSIFT | Densefuse | ECNN  | Proposed |
|-------------|-------|-------|-------|-----------|-------|----------|
| \( SSIM_a \) | 0.8462| 0.8428| 0.8428| 0.8669    | 0.8403| 0.8770   |
| \( Q^{AB/F} \) | 0.6049| 0.577 | 0.4923| 0.5594    | 0.4888| 0.660    |
| \( Q_{TE} \)  | 0.3964| 0.3972| 0.4029| 0.3812    | 0.4025| 0.4002   |
| \( Q_{MI} \)  | 0.7631| 0.7816| 0.7655| 0.8298    | 0.7668| 0.8560   |

We not only make subjective judgments on the experimental results, but also objectively evaluate them. Because the calculation of the quality evaluation index requires that the size of the fused image and the source image is the same, we first compress the fused image to the size of the source image in equal proportion, and then carry out the subsequent quality evaluation calculation. Table 1 shows the average values of 4 different quality evaluation metrics for 20 pairs of images in the “Lytro” dataset. The values in the Table 1 manifest that the proposed algorithm has higher \( SSIM_a \), \( Q^{AB/F} \) and \( Q_{MI} \) values
than other algorithms. In addition, although $Q_{TE}$ is not the highest, but it also ranks in the top three, which indicates that the optimized network performs well in the correlation between the source image and the fused image.

5. Conclusion
The optimized network has the following advantages. Firstly, more detailed features are extracted by dilated convolution combination of different scales. Secondly, the loss function is designed to optimize the image fusion network, so as to improve the fusion effect of the network and get the fusion image with more texture details. Thirdly, sub-pixel convolution is introduced to get super-resolution fusion image. The experimental results validated that the optimized method is superior to the existing algorithms in both of subjective and objective assessments. In the future, we will try to extend it to other image fusion applications, such as multi exposure image fusion and medical image fusion.

Acknowledgments
This work was partially funded by the National Natural Science Foundation of China, grant number 61976063

References
[1] Y. Liu, S. Liu, Z. Wang, Multi-focus image fusion with dense SIFT[J], Inf. Fusion 23 (2015) 139–155.
[2] S. Li, X. Kang, J. Hu, Image fusion with guided filtering[J], IEEE Trans. Image process. 22 (7) (2013) 2864–2875.
[3] H. Li, Y. Chai, H. Yin, et al., Multi-focus image fusion and denoising scheme based on homogeneity similarity[J], Optics Commun. 285 (2) (2012) 91–100.
[4] B. Yu, B. Jia, L. Ding, Z. Cai, Q. Wu, R. Law, J. Huang, L. Song, S. Fu, Hybrid dual-tree complex wavelet transform and support vector machine for digital multi-focus image fusion, Neuro computing 182 (2016) 1–9.
[5] Y. Liu, S. Liu, Z. Wang, A general framework for image fusion based on multi-scale transform and sparse representation[J], Information Fusion 24 (1) (2015) 147–164.
[6] Y. Liu, X. Chen, H. Peng, et al., Multi-focus image fusion with a deep convolutional neural network[J], Inf. Fusion 36 (2017) 191–207.
[7] H. Li, X.-J. Wu, Densefuse: a fusion approach to infrared and visible images[J], IEEE Trans.ImageProcess.28(5)(2019) 2626 14–23.
[8] Prabhakar K R, Srikar V S, Babu R V. DeepFuse: A Deep Unsupervised Approach for Exposure Fusion with Extreme Exposure Image Pairs[C]/2017 IEEE International Conference on Computer Vision(ICCV). IEEE, 2017: 4724-4732.
[9] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network[J]. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1874–1883, 2016.
[10] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution[J], In: European Conference on Computer Vision, Springer, 2016, pp. 694–711.
[11] 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.
[12] M. Amin-Naji, A. Aghagolzadeh, M. Ezoji, Ensemble of CNN for multi-focus image fusion[J], Inf. Fusion 51 (2019) 201–214 .
[13] Xydeas C S, Petrovic V. Objective image fusion performance measure[J]. Electronics letters, 2000, 36(4):308-309.
[14] Z. Liu, E. Blasch, Z. Xue, et al., Objective assessment of multiresolution image fusion algorithms for context enhancement in night vision: a comparative study[J], IEEE Trans. Pattern Anal. Mach. Intell. 34 (1) (2011) 94–109.
[15] S. Li, X. Kang, J. Hu, Image fusion with guided filtering[J], IEEE Trans. Image process. 22 (7) (2013) 2864–2875.