DTMNet: A Discrete Tchebichef Moments-based Deep Neural Network for Multi-focus Image Fusion

Bin Xiao Haifeng Wu Xiuli Bi*
Chongqing University of Posts and Telecommunications
Chongqing, China
xiaobin@cqupt.edu.cn S190231138@stu.cqupt.edu.cn bixl@cqupt.edu.cn

Abstract

Compared with traditional methods, the deep learning-based multi-focus image fusion methods can effectively improve the performance of image fusion tasks. However, the existing deep learning-based methods encounter a common issue of a large number of parameters, which leads to the deep learning models with high time complexity and low fusion efficiency. To address this issue, we propose a novel discrete Tchebichef moment-based Deep neural network, termed as DTMNet, for multi-focus image fusion. The proposed DTMNet is an end-to-end deep neural network with only one convolutional layer and three fully connected layers. The convolutional layer is fixed with DTM coefficients (DTMConv) to extract high/low-frequency information without learning parameters effectively. The three fully connected layers have learnable parameters for feature classification. Therefore, the proposed DTMNet for multi-focus image fusion has a small number of parameters (0.01M paras vs. 4.93M paras of regular CNN) and high computational efficiency (0.32s vs. 79.09s by regular CNN to fuse an image). In addition, a large-scale multi-focus image dataset is synthesized for training and verifying the deep learning model. Experimental results on three public datasets demonstrate that the proposed method is competitive with or even outperforms the state-of-the-art multi-focus image fusion methods in terms of subjective visual perception and objective evaluation metrics.

1. Introduction

Due to the limitation of imaging devices, it is difficult to capture an image where all objects are in focus. However, all-in-focus images are often required as input for specific computer vision tasks, such as localization, detection, and segmentation tasks [4]. A common method to solve the issues above is the multi-focus image fusion (MFIF) technique, which aims at obtaining an all-in-focus image by combining two or more images taken with diverse focal lengths [18]. Fig. 1 shows an example of the fused image obtained from two source images with different focal lengths. According to the fusion strategy, the existing MFIF methods can be roughly divided into two categories: traditional MFIF methods and deep learning-based MFIF methods.

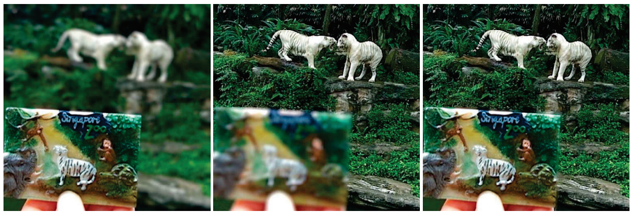

(a) Source image A (b) Source image B (c) Fused image

Fig. 1. An example of MFIF with two source images with different focal lengths. (a) Focus on foreground, (b) Focus on background and (c) All-in-focus.

The traditional MFIF methods can be further divided into transform domain-based methods and spatial domain-based methods. The transform domain-based methods firstly decompose the source images into multiple coefficients, and these different coefficients are then fused by following certain fusion rules to get the fused coefficients. Finally, the fused image is obtained by inverse transformation of the fused coefficients. Typical transform domain-based methods include gradient pyramid (GP) method [17] and discrete cosine transform (DCT) method [1]. Although the transform domain-based methods have high noise robustness and are easy to implement, the fused images always produce unreal results in brightness and color due to the imperfection of transformations and handcrafted features. As the name suggests, the source images are fused in the spatial domain for the spatial domain-based methods, i.e., using some spatial features of images [11]. Typical spatial domain-based methods include image matting(IM) method [8], guided filtering (GF) method [7] and density-SIFT (DSIFT) method [10]. Due to the influence of defocus and artificially de-

* Corresponding Author

signed focus measurement, the source image is blurred around the boundary between the focused/defocused area. Consequently, the fusion results of these methods are often blurred around the boundary [13].

The deep learning-based methods have become a very active direction in the field of MFIF in recent years. Liu *et al.* [9] first introduced convolutional neural network (CNN) into the field of MFIF. In their method, the activity level measurement and fusion rule, which are two crucial issues in the image fusion process, can be jointly generated by learning the CNN model, avoiding the manual design of focus measurement and fusion rules. Tang *et al.* [18] proposed a pixel-wise convolutional neural network (p-CNN) that can recognize the focused and defocused pixels in source images from its neighborhood information for MFIF. Some methods begin to use the convolutional layer to replace the fully connected layer because the fully connected layer consumes many storage resources. For example, Guo *et al.* [3] proposed an MFIF method based on a fully convolutional network (FCN). Although the existing deep learning-based MFIF methods can achieve good fusion performance, they usually improve the fusion performance by increasing the network depth or width, which will also increase the computational burden and the requirements of hardware, thereby reducing the efficiency of image fusion.

In this paper, in order to obtain high-quality fusion images with low time complexity, we propose a novel discrete Tchebichef moments-based deep neural network named DTMNet for MFIF. The contributions of this paper are summarized as follows:

- It is the first time that the image moments and deep learning technologies are combined to propose the lightweight end-to-end deep neural network, i.e., DTMNet for multi-focus image fusion.
- A DTM fixed convolution (DTMConv) is proposed, which can effectively extract the high/low-frequency information of the image and enhance the deep learning model's feature learning ability.
- In the proposed DTMNet, only the low-order Tchebichef polynomial coefficients and the 1×1 convolutional layer instead of fully connected layers are introduced to further reduce the parameters and improve the performance of the network.

The rest of this paper is organized as follows. In Section 2, the discrete Tchebichef moments (DTMs) is briefly introduced. In Section 3, we present the details of the proposed DTMNet. The extensive experiments conducted to evaluate the proposed method are presented in Section 4, and the conclusion is drawn in Section 5.

2. Discrete Tchebichef Moments

DTMs belong to a new image moments technology proposed in recent years. The kernel functions of DTMs are composed of discrete Tchebichef orthogonal polynomials with different orders and have the characteristic of fast iterative calculation [20]. Moreover, DTMs have the advantages of high de-correlation, no numerical approximation, and strong image reconstruction ability. They have been widely used in image analysis, recognition, and compression. In actual implementation, the DTMs are computed with the kernel matrix [20], as is shown in Fig. 2.

$$\boldsymbol{E} = \begin{bmatrix} -0.3333 & 0.3333 & 0.3333 & 0.3333 & 0.3333 & 0.3333 & 0.3333 & 0.3333 & 0.3333 \\ -0.5164 & -0.3873 & -0.2582 & -0.1291 & 0 & 0.1281 & 0.2582 & 0.3873 & 0.5164 \\ 0.5318 & 0.1330 & -0.1519 & -0.3229 & 0 & -0.3229 & -0.1519 & 0.1330 & 0.5318 \\ -0.4449 & 0.2225 & 0.4132 & 0.2860 & 0 & -0.2860 & -0.4132 & -0.2225 & 0.4449 \\ 0.3219 & -0.4693 & -0.2458 & 0.2011 & 0 & 0.2011 & -0.2458 & -0.4693 & 0.3129 \\ -0.1849 & 0.5085 & -0.1849 & -0.4610 & 0 & 0.4610 & 0.1849 & -0.5085 & 0.1849 \\ 0.0899 & -0.3820 & 0.4944 & 0.0225 & 0 & 0.0225 & 0.4944 & -0.3820 & 0.0899 \\ -0.0341 & 0.2048 & -0.4780 & 0.4780 & 0 & -0.4780 & 0.4780 & -0.2048 & 0.0341 \\ 0.0088 & -0.0705 & 0.2468 & -0.4936 & 0 & -0.4936 & 0.2468 & -0.0705 & 0.0088 \end{bmatrix}$$

Fig. 2. A typical 9×9 kernel matrix of DTMs, in which each row represents a DTM polynomial with different orders.

2.1. DTMs as Correlation

Correlation is a similarity measure between two functions. The correlation $\mathbb{R}_{gt}(a,b)$ of two discrete functions $g(x,y) \in R^{M\times N}$ and $t(x,y) \in \mathrm{R}^{M\times N}$ are defined by [25] as

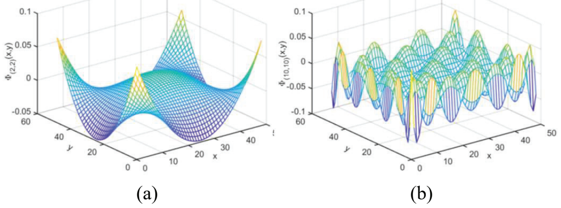
$$\mathbb{R}_{gt}(a,b) = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} g(x,y)t(x-a,y-b). \quad (1)$$

Given a digitalized image $f(x,y)$ with the size of $M \times N$, the $(m+n)^{th}$ order of DTM of image is defined by [14] as

$$K_{m,n} = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} \tilde{k}_m(x;M)\tilde{k}_n(y;N)f(x,y), \quad (2)$$

where $m = 0,1,\ldots,M-1$, $n = 0,1,\ldots,N-1$ and $\tilde{k}_m(x;M)\tilde{k}_n(y;N)$ is the kernel function of DTMs. According to Eq. (1) and (2), we can conclude that the DTMs of image $f(x,y)$, i.e., $(K_{m,n})$ is actually the correlation between image $f(x,y)$ and the DTM kernel function $\tilde{k}_m(x;M)\tilde{k}_n(y;N)$. In other words, DTMs measure the similarity between image and the kernel functions of DTMs.

According to the frequency distribution of DTMs' kernel functions, the DTMs with different orders $(K_{m,n})$ measure different spatial frequency components of an image. The lower order DTMs measure the low spatial frequency components of the image, while the higher-order DTMs measure the high spatial frequency components of the image. This can also be verified by observing the plots of the kernel functions of DTMs. For simplicity, we denote $\Phi_{m,n}(\mathrm{x, y})$ =$\tilde{k}_m(x;M)\tilde{k}_n(y;N)$ as the kernel function of DTMs, and Fig. 3 shows the plots of DTMs' kernel func-



(a) (b)

Fig. 3. The plots of DTMs' kernel function $\Phi_{m,n}$. (a) $\Phi_{2,2}$ and (b) $\Phi_{10,10}$. The x and y in the figure represent the pixel position on the image.

tions with different orders. As shown in this figure, the value of kernel function $\Phi_{2,2}$ changes smoothly, indicating that $K_{2,2}$ measures the low spatial frequency component of the image. On the contrary, the value of kernel function $\Phi_{10,10}$ changes drastically, indicating that $K_{10,10}$ measures the high spatial frequency component of the image.

2.2. Focus Measurement Based on DTMs

The change of image focusing degree is mainly reflected in the change of high spatial frequency components of an image. Combining the conclusion from subsection 2.1, the focus measurement based on DTMs was defined by [26] as

$$FM_{\mathrm{K}} = \frac{||H(\tilde{f};p)||}{||L(\tilde{f};p)||}, \tag{3}$$

where $||F|| = {F_1}^2 + {F_2}^2 + \cdots + {F_n}^2$ denotes the energy of F, $p = m + n$ is the order of DTMs, $\tilde{f}$ is the normalized image block with the size of $b \times b$, which is defined as

$$\tilde{f}(x,y) = \frac{f(x,y)}{\sqrt{\sum_{x=0}^{b-1} \sum_{y=0}^{b-1} \left[f(x,y)\right]^2}}, \tag{4}$$

and $\tilde{f}$ satisfies the following property:

$$\sum\nolimits_{x=0}^{b-1} \sum\nolimits_{y=0}^{b-1} \left[\tilde{f}(x,y)\right]^2 = 1. \tag{5}$$

$L(\tilde{f};p)$ and $H(\tilde{f};p)$ denote the sets of low-order and high-order DTMs,

$$\begin{aligned} &L(\tilde{f};p) = \{K_{m,n} | m+n \leq p\}, \\ &H(\tilde{f};p) = \ \{K_{m,n} | m+n > p\}, \\ &M + N - 2 \geq p \geq 0. \end{aligned} \tag{6}$$

Combining Eqs. (4), (5), (6) and the Parseval theorem, $L(\tilde{f};p)$ and $H(\tilde{f};p)$ satisfy the following property:

$$\begin{aligned} &||L(\tilde{f};p)|| + ||H(\tilde{f};p)|| = ||\tilde{f}|| = 1, \\ &||L(\tilde{f};p)|| = 1 - ||H(\tilde{f};p). \end{aligned} \tag{7}$$

Thus, the focus measure based on DTMs defined in Eq. (3) can be simplified as

$$FM_{\mathrm{K}} = \frac{||H(\tilde{f};p)||}{||L(\tilde{f};p)||} = \frac{1-||L(\tilde{f};p)||}{||L(\tilde{f};p)||}. \tag{8}$$

It can be seen from Eq. (8) that only the set of low-order DTMs can calculate the focus measurement FM_{K}. Moreover, when the value of p is small, the focus measurement's computational complexity can be greatly reduced. Therefore, only the low-frequency components of an image captured by the set of low-order DTMs can be used to measure the image focusing degree.

3. The Proposed DTMNet for MFIF

The framework of MFIF by the proposed DTMNet is shown in Fig. 4. In this framework, we mainly consider the task of fusing two source images with different focal lengths, and the fusion of three or more source images can be straightforwardly extended based on this framework.

3.1. Basic Modules

1) DTMConv Block: The investigation on the correlation between image and DTMs' kernel functions provided in subsection 2.1 demonstrates that the high/low-order DTMs can effectively extract the high/low-frequency information of image, and the investigation in subsection 2.2 shows that only the low-frequency components of an image captured by the set of low-order DTMs can be used to measure the image focusing degree. Moreover, the image information stored in each moment is independent and the information redundancy between the moments is minimal. Thus, we design the DTMConv block (shown in Fig. 5), which is composed of a normalization layer (Norm) and a DTM convolution layer (DTMConv), to extract the low-frequency feature.

Firstly, the normalization layer is used to preprocess the input image to enhance the non-numerical approximation property of DTMs. The normalization process is expressed as

$$\overline{f}(x,y) = \frac{f(x,y)}{\sqrt{\sum_{x=x-b/2}^{x+b/2} \sum_{y=y-b/2}^{y+b/2} \left[f(x,y)\right]^2}}, \tag{9}$$

where $f(x,y)$ represents the gray-scale image, and the size of the normalized image block is $b \times b$, which is consistent with the size of the convolution kernel of DTMConv.

Secondly, the DTM convolutional layer is composed of weights and bias, which are used to extract low-frequency features of the image. The weights of filters in DTMConv can be expressed as

$$W_n^p = (E_i)^T E_j, \left(i+j=p,\ p \leq 5,\ n = \sum_{t=0}^{i+j} (t+1) - j\right), \tag{10}$$

where E is the kernel matrix of DTMs with the size of $h \times h$, $(E_i)^T$ is the $h \times 1$ column vector obtained by transposing the $(i+1)^{th}$ row of E, E_j is the $(j+1)^{th}$ row of E, $p = i + j$ is the order of DTMs, n is the n^{th} convolution kernel in DTMConv. Therefore, W_n^p represents the weight of the n^{th} convolution kernel, which is obtained by the coefficients of

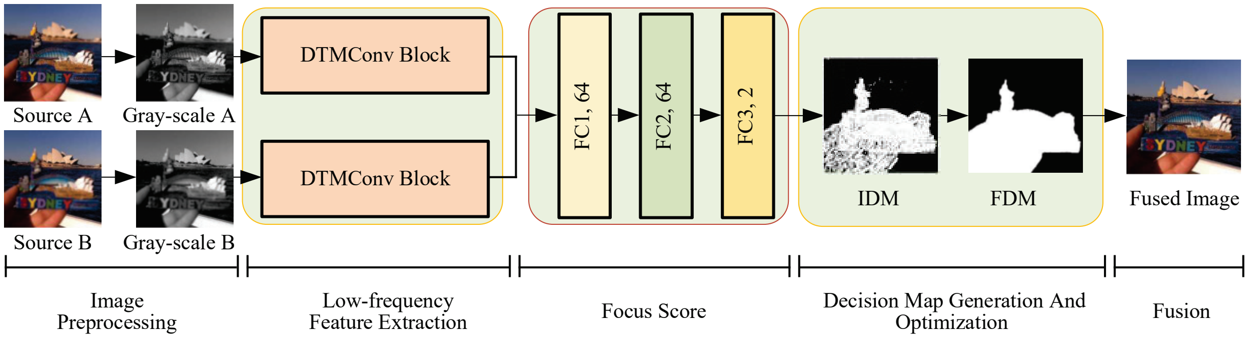


Fig. 4. The proposed MFIF algorithm using DTMNet.

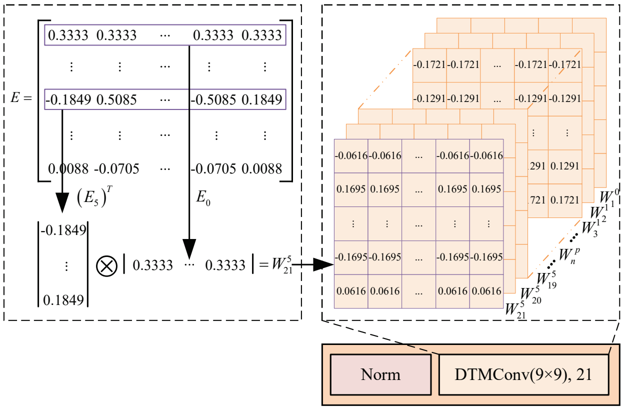


Fig. 5. The detail description of the DTMConv Block. E_j represents the $(j+1)^{th}$ row of the kernel function E, $(E_i)^T$ represents the transpose operation on E_i, and $\otimes$ represents matrix multiplication.

polynomial in DTMs with order p.

In this paper, the kernel matrix E of DTMs is a 9×9 matrix presented in Fig. 2. As shown in Fig. 5, when $p = 5$, i and j take 5 and 0 respectively, i.e., the transpose of the 6^{th} row ($(E_5)^T$) of the DTMs' kernel matrix E is multiplied by the 1^{th} row ($(E_0)^T$) of the DTMs' kernel matrix E to obtain a 9×9 matrix W_{21}^5, which is used as the weight of the 21^{th} convolution kernel of the DTMConv. According to Eq. (10), when $p \leq 5$, the available weights are 21. Therefore, the low 5-order polynomials in DTMs can get 21 convolution kernels. In addition, the bias of the DTMConv are set to 0.

2) Fully Connected Layer: The detailed description of the fully connected (FC) layer is shown in Fig. 6. In our experiment, we treat the MFIF task as a binary classification task. We make a focus evaluation for each pixel of the image to determine whether it is focused or de-focused. The general classification method is through the fully connected layer, but our experiment needs to meet the following three requirements: First, sharing parameters. Second, keeping the spatial structure of the image and the size of the feature map unchanged during the classification process. Third, the number of parameters should be as small as possible.

In this paper, we introduce three 1×1 convolution layers to replace the fully connected layer. The number of filters in the three convolutional layers is 64, 64, and 2, respectively. It is noted that we also add the ReLU nonlinear activation function after the first and second convolutional layers to increase the nonlinearity of the network and enable our network to express more complex features. Moreover, in order to classify a two-class focus evaluation corresponds to each pixel, the softmax activation function is added after the last convolutional layer. Finally, a two-channel feature map can be obtained after the focus score module, which is the focus score of each pixel corresponds to the two source images.

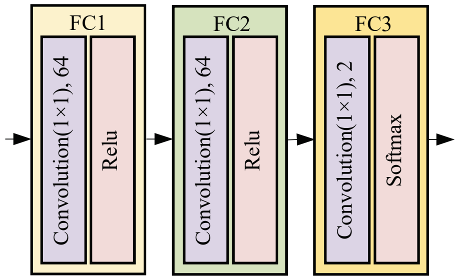
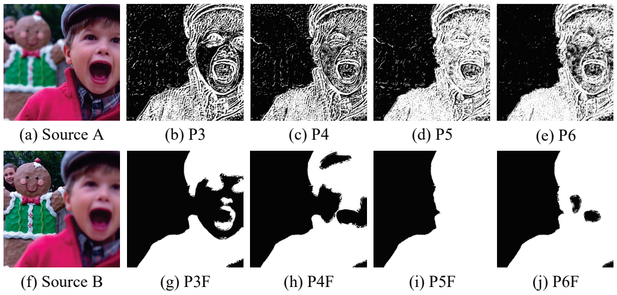


Fig. 6. Detail description of fully connected layers.

3.2. Effect Order P

Generally, higher-order DTMs can capture more high-frequency content of images. However, it is confirmed in [25] that with the increase of DTMs order, DTMs implies better discriminating power between various degrees of blurring, but it also means that the moments are more susceptible to the effect of noise.

In order to have a suitable setting for the order of DTMs, we verify it on the Lytro dataset, as shown in Fig. 7. Pn represents p = n, and PnF represents the final decision map generated by Pn. Fig.7 (a) and (f) are a pair of multi-focus source images on the Lytro dataset. Fig. 7 (b)-(e) represent the initial decision maps generated when p values are 3, 4, 5, and 6 respectively. Fig. 7 (g)-(j) are the final decision maps optimized by CRF for the initial decision maps of corresponding orders. The results show that when p = 5,



(a) Source A (b) P3 (c) P4 (d) P5 (e) P6

(f) Source B (g) P3F (h) P4F (i) P5F (j) P6F

Fig. 7. The decision map generated by DTMs of different orders.

Method	Q_{MI}	Q_G	Q_Y	Q_{CB}	Parameter(M)	Time(s)
P3F	1.0849	0.6975	0.9817	0.7512	0.0065	0.3131
P4F	1.0939	0.7058	0.984	0.7685	0.0076	0.3201
P5F	**1.1033**	**0.7134**	**0.9888**	**0.7861**	0.0089	0.3371
P6F	1.0990	0.7114	0.9856	0.7819	0.0103	0.3551

Table 1. Comparison of the fused images generated by different orders in terms of quantitative indicators, parameters and calculation time. For Q_{MI} [5], Q_G [23], Q_Y [24] and Q_{CB} [2], the larger values indicate better results. The best result is in bold.

the decision map is the best. In addition, Table 1 also shows the comparison of the fusion images generated by the final decision maps of different orders in terms of objective indicators, the parameters, and the calculation time. It is found that with the increase of order, the number of parameters and calculation time of DTMNet increase positively, but the quantitative results reach the peak at p = 5. Therefore, p = 5 is the optimal order of DTMS.

3.3. Fusion Details

In this paper, we introduce the DTMNet architecture for MFIF. As shown in Fig. 4, the fusion process consists of five steps: image preprocessing, low-frequency feature extraction, focus score, decision map generation and optimization, and final image fusion. Firstly, the source images A and B are converted into gray-scale images. Secondly, the gray-scale images are fed into the DTMConv block to extract the low-frequency components of the images. Thirdly, the low-frequency feature maps are first concatenated and fed into the focus score module for feature classification and focus measurement. Then two-channel feature maps of the same size as the source image are obtained, in which each value represents the focus score of the corresponding pixel in the two source images. Fourthly, the two-channel focus maps are used to generate a binary image with the same size as the source image, i.e., the initial decision map (IDM). Then, we employ the Conditional Random Field (CRF) to optimize IDM so that the final decision map (FDM, D) is obtained for final MFIF. Finally, the fused image F_{fusion} is obtained using the pixel-wise weighted-average strategy as

$$F_{fusion}(x,y) = D(x,y) \odot A(x,y) + (1 - D(x,y)) \odot B(x,y), \tag{11}$$

where $\odot$ denotes the dot product.

3.4. Loss function

In the proposed DTMNet, MFIF is regarded as a binary classification task with a two-channel output. Compared with single-channel output, two-channel output considers the relationship between each pair of input source images. Therefore, we introduce a binary cross-entropy loss function to fully exploit each pair of source images' complementary correlation.

In the training process, the binary cross-entropy loss function calculates the error between the focus maps f output by DTMNet and the ground truth focus maps $\hat{f}$, and then backpropagates to optimize the model parameters, which is defined as

$$L_{loss}(f,\hat{f}) = -\frac{1}{NC}\sum_{i=1}^{N}\sum_{j=1}^{C}[\hat{f}_j^{(i)}\log(f_j^{(i)}) + (1 - f_j^{(i)})\log(1 - \hat{f}_j^{(i)})], \tag{12}$$

where N is the batch size, C is the number of channel.

4. Experiment and Discussion

4.1. Experimental Settings

1) Datasets setting: The training and validation datasets are generated from the Benchmark dataset [19], which contains 10569 all-in-focus images and 10569 separated mask images for the target. We adopted a Gaussian filter with a standard deviation of 3 and a window of 7 × 7 to continuously smooth each image, and finally obtained 10569 all-in-blur images. Then, we generated a pair of multi-focus images I_a and I_b based on the existing all-in-focus image I_{clear}, the all-in-blur image I_{blur} and the mask M_{clear}:

$$\begin{aligned} I_a &= I_{clear} \odot M_{clear} + I_{blur} \odot (1 - M_{clear}), \\ I_b &= I_{clear} \odot (1 - M_{clear}) + I_{blur} \odot M_{clear}, \end{aligned} \tag{13}$$

where $\odot$ denotes the dot product. Finally, 10569 pairs of synthesized multi-focus images and the corresponding ground-truth mask are obtained. To effectively train the proposed DTMNet model, we divide the generated dataset into training dataset and validation dataset at a ratio of 9:1.

The testing datasets come from three public datasets: Lytro dataset [15], Nature dataset [27], MFFW dataset [22]. The Lytro dataset is widely used to test the performance of MFIF algorithms and contains a total of 20 pairs of multi-focus images with the size of 520×520. The Nature dataset contains a total of 16 pairs of multi-focus images with variable sizes, which are more difficult to distinguish between focus and defocus boundaries(FDB) than the Lytro dataset. The MFFW dataset contains a total of 13 pairs of multi-focus images with variable sizes, which considers the effect of defocus propagation.

2) Training setting: We implemented the training process by PyTorch [16]. The model ran on the NVIDIA Tesla V100 GPU of memory size 16GB with CUDA version 10.1 and CUDNN version 6.0. It costs around 3 GB of GPU memory to train our model. The model was trained using the Adam optimizer with the mini-batch composed of 32 training image pairs, which were selected from the reshuffled training data. During the training process, the images in the training dataset are resized to 256 × 256. The learning rate is initialized to 0.0001 and decreased to 0.00001 at the 20^{th} iteration by setting the weight decay factor to 0.1. We set the maximum epoch number to 60 and validated it after every two training epochs. In addition, to enhance the generalization ability and the fusion effectiveness of the model, data enhancement technology was applied, including horizontal and vertical flipping.

4.2. Subjective Evaluation and Analysis

In this section, the proposed DTMNet method is compared with nine state-of-the-art MFIF methods. The transform domain-based methods: DCT [1]. The spatial domain-based methods: IM [8], GF [7] and DSIFT [10]. The deep learning-based methods: CNN [9], p-CNN [18], DRPL [6], MMF-Net [13] and GEU-Net [21]. For a fair and comprehensive comparison, the source codes of all comparison methods are provided by the corresponding authors, and the quantitative evaluation is carried out on the three public datasets including Lytro, Nature and MFFW.

1) Experiments on the Lytro dataset: Fig. 8 shows the effectiveness of the proposed DTMNet on the conventional dataset. In the enlarged image, the club in source image A (Fig. 8 (a)) is focused and smooth, and four focused and complete baseballs are presented in source image B (Fig. 8 (b)). The brightness of the transform domain method (Fig. 8 (c)) is not realistic, and the image details at the boundary are destroyed. The incomplete and out-of-focus baseballs both appear in the spatial domain methods (Fig. 8 (d)-(f)) and the deep learning methods (Fig. 8 (g) and (i)). The out-of-focus area also appears in the club in Fig. 8 (h). In contrast, the proposed DTMNet method (Fig. 8 (l)) obtains a fused image with the smooth club and complete baseballs.

We also employed the difference image with the source image B to clearly show the comparison of different methods because the visual results are not always easy to distinguish. The difference map can be obtained as

$$DifferenceMap = |Fused - SourceB| \,. \tag{14}$$

Because the focus of the source image B is in the background and it is out of focus in the foreground, the difference image with a satisfactory fusion result would be all black in the background. Fig. 9 shows the difference image between the fused image in Fig. 8 and the source image B (Fig. 8(b)). The proposed DTMNet (Fig. 9 (l)) is smooth on the edge and black in the background, while other comparison methods (Fig. 9 (a)-(h)) have noise in the background.

2) Experiments on the Nature dataset: Fig. 10 shows the effectiveness of the proposed DTMNet method on the dataset with complex FDB. Fig. 11 shows the difference images. In the enlarged image, the apical part of a petal is in focus in source A (Fig. 10 (a)) and out of focus in source B (Fig. 10 (b)), which indicates that the difference image should not have an apical part. Unclear regions remain in Fig. 10 (c)-(e), (g) and (k) because the apical part exists in Fig. 11 (a)-(c), (e) and (i). The difference image in Fig. 11 (f) is not smooth at the boundary, indicating that p-CNN has poor robustness to the complex boundary. Although the difference image in Fig. 11 (h) is accurate but the color is changed, indicating that the content of the fused image obtained by the MMF-Net method has been changed. In contrast, the proposed DTMNet produces a clear result in Fig. 10 (l) even in such a difficult situation due to the lack of apical part in the difference image in Fig. 11 (j),

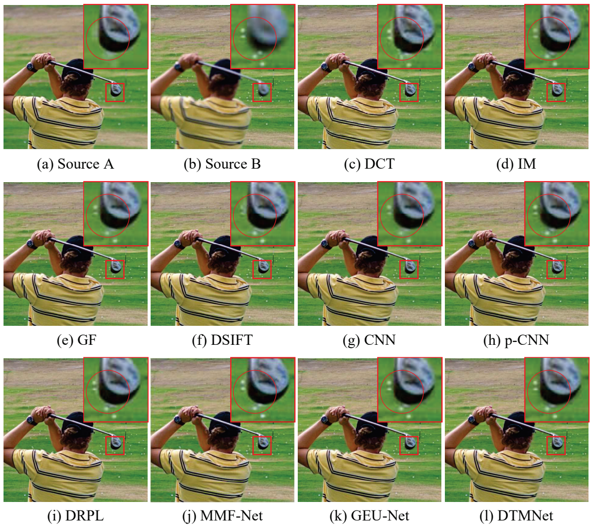


Fig. 8. Experimental results on a pair of source images from the Lytro dataset [15].

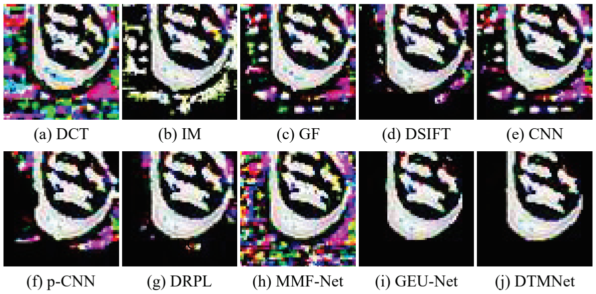


Fig. 9. The difference image on Lytro dataset [15].

3) Experiments on the MFFW dataset: Fig. 12 shows the effectiveness of the proposed DTMNet method on the dataset with defocus effect. Fig. 13 shows the difference images. In the enlarged image, the cup in source A (Fig. 12 (a)) is in focus, and the branches in source B (Fig. 12 (b)) are relatively in focus due to the defocus effect. In Fig. 13, a lot of noise remains above the cup, indicating that the fusion results generated by these methods are blurred in the

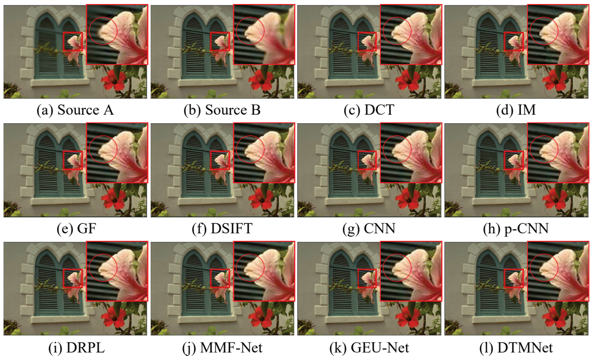

(a) Source A (b) Source B (c) DCT (d) IM
(e) GF (f) DSIFT (g) CNN (h) p-CNN
(i) DRPL (j) MMF-Net (k) GEU-Net (l) DTMNet

Fig. 10. Experimental results on a pair of source images from the Nature dataset [27].

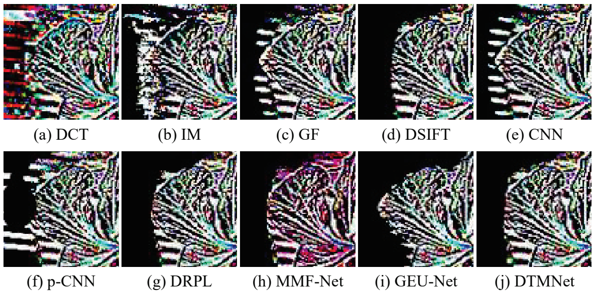

(a) DCT (b) IM (c) GF (d) DSIFT (e) CNN
(f) p-CNN (g) DRPL (h) MMF-Net (i) GEU-Net (j) DTMNet

Fig. 11. The difference image on Nature dataset [27].

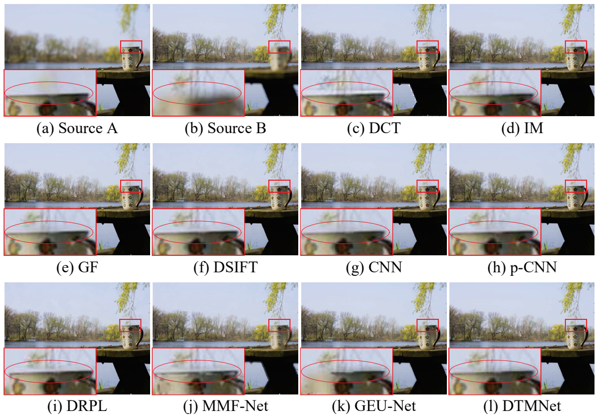

(a) Source A (b) Source B (c) DCT (d) IM
(e) GF (f) DSIFT (g) CNN (h) p-CNN
(i) DRPL (j) MMF-Net (k) GEU-Net (l) DTMNet

Fig. 12. Experimental results on a pair of source images from the MFFW dataset [22].

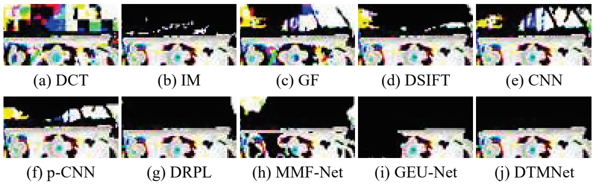

(a) DCT (b) IM (c) GF (d) DSIFT (e) CNN
(f) p-CNN (g) DRPL (h) MMF-Net (i) GEU-Net (j) DTMNet

Fig. 13. The difference image on MFFW dataset [22].

corresponding area, as shown in Fig. 12 (c)-(h) and (j). In addition, the incomplete difference image exists in Fig. 13 (i), resulting in severe defocus as shown in Fig. 12 (k). In contrast, the difference image is complete and noise-free in Fig. 13 (l), which indicates that the proposed method has good robustness to the defocus effect.

4) Experiments on other images: Fig. 14 shows the difference images of the remaining images. The noise almost exists in all images in Fig. 14 (b) and (i), and the color of some images is changed, which shows that the fused results generated by DCT and MMF-Net have changed the image components. In Fig. 14 (c)-(f) and (h), the blur exists at the boundary of the grid in the first row, and the noise remains in the last two rows. In addition, the grid in the first row of Fig. 14(g) and (j) is interrupted. On the contrary, the proposed DTMNet achieves smooth and complete fusion results in all images.

4.3. Objective Evaluation and Analysis

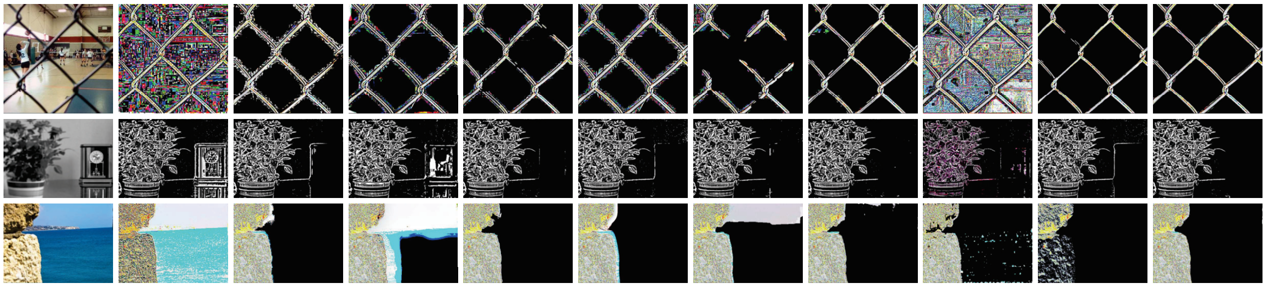
To further illustrate the effectiveness of the proposed DTMNet for MFIF, we verified the images fused by different methods from the objective evaluation. In [12], the evaluation metrics used for MFIF are roughly classified into four categories: information theory-based metrics, image feature-based metrics, image structural similarity-based metrics and human perception inspired fusion metrics. In this paper, we select a typical evaluation metric from each type of objective evaluation metrics to verify the performance of the fusion methods, i.e., normalized mutual information (Q_{MI}) [5], gradient-based fusion performance (Q_G) [23], structural similarity (SSIM, Q_Y) [24] and Chen-blum metric (Q_{CB}) [2].

The objective evaluation on the three public datasets are shown in Table 2. For the Lytro dataset, the values of the proposed DTMNet method are higher than those of other comparison methods on three metrics, i.e., Q_{MI}, Q_G and Q_Y. For the Nature and MFFW datasets, the values of the proposed DTMNet method are the highest in all metrics among all methods.

4.4. Comparison in Parameters and Computational Time

Since the traditional method has no learnable parameters, we only compared the deep learning-based methods from the aspect of learnable parameters, as shown in Table 3. We can find in this table the proposed DTMNet has the least parameters among all the comparison methods, with only 0.01M parameters, which is 1/500 times that of CNN.

In addition, Table 3 also lists the average running time for each method to generate a fusion image on the Lytro dataset. The comparison is based on a platform with an Intel Core i5-8300H CPU, an NVIDIA Geforce GTX 1060 GPU, and an 8GB RAM. Generally, the traditional methods have higher efficiency than the deep learning methods. However, we can find that the DTMNet method is much faster than all comparison methods, including traditional methods and deep learning-based methods. Therefore, thanks to a very small number of parameters, the proposed DTMNet can be used in some real-time MFIF applications.



(a) Source B (b) DCT (c) IM (d) GF (e) DSIFT (f) CNN (g) p-CNN (h) DRPL (i) MMF-Net (j) GEU-Net (k) DTMNet

Fig. 14. The difference images of the remaining images on the three datasets.

Method	Lytro dataset				Nature dataset				MFFW dataset			
	Q_{MI}	Q_G	Q_Y	Q_{CB}	Q_{MI}	Q_G	Q_Y	Q_{CB}	Q_{MI}	Q_G	Q_Y	Q_{CB}
DCT [1]	0.8357	0.6110	0.9146	0.6738	0.9377	0.5974	0.8894	0.6811	0.7985	0.5646	0.8755	0.6291
IM [8]	1.1419	0.7174	0.9787	0.7952	1.2260	0.7111	0.9747	0.7884	1.0932	0.6893	0.9652	0.7364
GF [7]	1.0980	0.7204	0.9817	0.7975	1.1397	0.7019	0.9640	0.7767	0.9594	0.6779	0.9652	0.7364
DSIFT [10]	1.1876	0.7266	0.9877	**0.8093**	1.2565	0.7184	0.9797	0.8048	1.1426	0.6906	0.9404	0.7341
CNN [9]	1.1512	0.7250	0.9871	0.8084	1.2286	0.7180	0.9852	0.8040	1.0915	0.6818	0.9735	0.7415
p-CNN [18]	1.1773	0.7234	0.9860	0.8049	1.2406	0.7077	0.9707	0.7913	1.1338	0.6839	0.9625	0.7393
DRPL [6]	1.1895	0.7274	0.9867	0.8067	1.2482	0.7049	0.9614	0.7831	1.1462	0.6886	0.9312	0.7152
MMFNet [13]	0.9719	0.6576	0.9517	0.7508	1.0554	0.6340	0.8950	0.7039	0.9535	0.6001	0.8705	0.6410
GEU-Net [21]	1.1564	0.7140	0.9798	0.7858	1.2545	0.7126	0.9809	0.7910	1.1215	0.6971	0.9780	0.7494
DTMNet	**1.1928**	**0.7282**	**0.9886**	0.8089	**1.2692**	**0.7197**	**0.9862**	**0.8081**	**1.1853**	**0.7036**	**0.9834**	**0.7510**

Table 2. The quantitative comparison of different MFIF methods. For Q_{MI} [5], Q_G [23], Q_Y [24] and Q_{CB} [2], the larger values indicate better results. The best result is in bold.

Method	Traditional methods				Deep learning methods					
	DCT	IM	GF	DSIFT	CNN	p-CNN	DRPL	MMFNet	GEU-Net	DTMNet
Para(M)	-	-	-	-	4.93	0.31	1.07	5.22	2.16	**0.01**
Time(s)	0.49	16.94	2.74	3.36	79.09	3.52	3.55	8.92	1.65	**0.32**

Table 3. The number of parameters (Para) for the various models and the average time consumption to generate the fused image on the Lytro dataset.

Method	Conv	DCTConv	DTMConv
Q_{MI}	1.1862	1.1891	**1.1928**
Q_G	0.6259	0.6467	**0.7282**
Q_Y	0.9158	0.9383	**0.9886**
Q_{CB}	0.7182	0.7344	**0.8089**

Table 4. Ablation Study. We use the quantitative comparisons on the Lytro dataset with different training settings.

4.5. Ablation Study

Several ablation studies are conducted to show the effectiveness of the proposed DTMConv for feature extraction in Table 4. The experimental settings remain unchanged, and DTMConv is replaced with learnable convolution and discrete cosine convolution (DCTConv) respectively. The results show that the DTMConv has a stronger image representation ability than learnable convolution and DCT, which helps extract more abundant image information.

5. Conclusion

In this paper, a novel discrete Tchebichef moments-based convolutional neural network, termed as DTMNet, is proposed for MFIF. To the best of our knowledge, it is the first time that the image moment and deep learning technologies are combined to propose the lightweight end-to-end deep neural network, i.e., DTMNet for MFIF. The proposed DTMNet has 0.01M parameters only, and it can be used in some real-time MFIF applications. Extensive experiments on three public datasets show that the proposed DTMNet outperforms the state-of-the-art methods in terms of visual quality and objective assessment.

Acknowledgements

This work was supported in part by the National Key Research and Development Project under Grant 2019YFE0110800 and Grant 2016YFC1000307-3, in part by the National Natural Science Foundation of China under Grant 61976031 and Grant 61806032, in part by the National Major Scientific Research Instrument Development Project of China under Grant 62027827, in part by the Scientific and Technological Key Research Program of Chongqing Municipal Education Commission under Grant KJZD-K201800601.

References

[1] Liu Cao, Longxu Jin, Hongjiang Tao, Guoning Li, Zhuang Zhuang, and Yanfu Zhang. Multi-focus image fusion based on spatial frequency in discrete cosine transform domain. *IEEE signal processing letters*, 22(2):220–224, 2014.

[2] Yin Chen and Rick S Blum. A new automated quality assessment algorithm for image fusion. *Image and vision computing*, 27(10):1421–1432, 2009.

[3] Xiaopeng Guo, Rencan Nie, Jinde Cao, Dongming Zhou, and Wenhua Qian. Fully convolutional network-based multifocus image fusion. *Neural computation*, 30(7):1775–1800, 2018.

[4] Rania Hassen, Zhou Wang, and Magdy MA Salama. Objective quality assessment for multiexposure multifocus image fusion. *IEEE transactions on image processing*, 24(9):2712–2724, 2015.

[5] M Hossny, S Nahavandi, and D Creighton. Comments on'information measure for performance of image fusion'. *Electronics letters*, 44(18):1066–1067, 2008.

[6] Jinxing Li, Xiaobao Guo, Guangming Lu, Bob Zhang, Yong Xu, Feng Wu, and David Zhang. Drpl: Deep regression pair learning for multi-focus image fusion. *IEEE Transactions on Image Processing*, 29:4816–4831, 2020.

[7] Shutao Li, Xudong Kang, and Jianwen Hu. Image fusion with guided filtering. *IEEE Transactions on Image processing*, 22(7):2864–2875, 2013.

[8] Shutao Li, Xudong Kang, Jianwen Hu, and Bin Yang. Image matting for fusion of multi-focus images in dynamic scenes. *Information Fusion*, 14(2):147–162, 2013.

[9] Yu Liu, Xun Chen, Hu Peng, and Zengfu Wang. Multi-focus image fusion with a deep convolutional neural network. *Information Fusion*, 36:191–207, 2017.

[10] Yu Liu, Shuping Liu, and Zengfu Wang. Multi-focus image fusion with dense sift. *Information Fusion*, 23:139–155, 2015.

[11] Yu Liu, Lei Wang, Juan Cheng, Chang Li, and Xun Chen. Multi-focus image fusion: A survey of the state of the art. *Information Fusion*, 64:71–91, 2020.

[12] Zheng Liu, Erik Blasch, Zhiyun Xue, Jiying Zhao, Robert Laganiere, and Wei Wu. Objective assessment of multiresolution image fusion algorithms for context enhancement in night vision: a comparative study. *IEEE transactions on pattern analysis and machine intelligence*, 34(1):94–109, 2011.

[13] Haoyu Ma, Qingmin Liao, Juncheng Zhang, Shaojun Liu, and Jing-Hao Xue. An α-matte boundary defocus model-based cascaded network for multi-focus image fusion. *IEEE Transactions on Image Processing*, 29:8668–8679, 2020.

[14] R. Mukundan, S. H. Ong, and P. A. Lee. Image analysis by tchebichef moments. *IEEE Transactions on Image Processing*, 10(9):1357–1364, 2001.

[15] Mansour Nejati, Shadrokh Samavi, and Shahram Shirani. Multi-focus image fusion using dictionary-based sparse representation. *Information Fusion*, 25:72–84, 2015.

[16] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[17] Vladimir S Petrovic and Costas S Xydeas. Gradient-based multiresolution image fusion. *IEEE Transactions on Image processing*, 13(2):228–237, 2004.

[18] Han Tang, Bin Xiao, Weisheng Li, and Guoyin Wang. Pixel convolutional neural network for multi-focus image fusion. *Information Sciences*, 433:125–141, 2018.

[19] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2411–2418, 2013.

[20] Bin Xiao, Gang Lu, Yanhong Zhang, Weisheng Li, and Guoyin Wang. Lossless image compression based on integer discrete tchebichef transform. *Neurocomputing*, 214:587–593, 2016.

[21] Bin Xiao, Bocheng Xu, Xiuli Bi, and Weisheng Li. Global-feature encoding u-net (geu-net) for multi-focus image fusion. *IEEE Transactions on Image Processing*, 30:163–175, 2020.

[22] Shuang Xu, Xiaoli Wei, Chunxia Zhang, Junmin Liu, and Jiangshe Zhang. Mffw: A new dataset for multi-focus image fusion. *arXiv preprint arXiv:2002.04780*, 2020.

[23] CS Xydeas and Vladimir Petrovic. Objective image fusion performance measure. *Electronics letters*, 36(4):308–309, 2000.

[24] Cui Yang, Jian-Qi Zhang, Xiao-Rui Wang, and Xin Liu. A novel similarity based quality metric for image fusion. *Information Fusion*, 9(2):156–160, 2008.

[25] Pew Thian Yap and P Raveendran. Image focus measure based on chebyshev moments. *IEE Proceedings-Vision, Image and Signal Processing*, 151(2):128–136, 2004.

[26] P. T. Yap and P. Raveendran. Image focus measure based on chebyshev moments. *IEE Proceedings - Vision, Image and Signal Processing*, 151(2):128–136, 2004.

[27] Yu Zhang, Xiangzhi Bai, and Tao Wang. Boundary finding based multi-focus image fusion through multi-scale morphological focus-measure. *Information fusion*, 35:81–101, 2017.