IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 8, 2018, accepted January 28, 2019, date of publication January 31, 2019, date of current version February 22, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2896655

Focus Measure for Synthetic Aperture Imaging Using a Deep Convolutional Network

ZHAO PEI[1,2,3], LI HUANG[2], YANNING ZHANG[4], (Senior Member, IEEE), MIAO MA[2], YALI PENG[2], AND YEE-HONG YANG[3], (Senior Member, IEEE)

[1] Key Laboratory of Modern Teaching Technology, Ministry of Education, Xi'an 710062, China
[2] School of Computer Science, Shaanxi Normal University, Xi'an 710119, China
[3] Department of Computing Science, University of Alberta, Edmonton, AB T6G2E8, Canada
[4] School of Computer Science, Northwestern Polytechnical University, Xi'an 710129, China

Corresponding author: Miao Ma (mmthp@snnu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61501286, Grant 61877038, Grant 61672333, Grant 61671385, and Grant 61402274, in part by the Key Research and Development Program in Shaanxi Province of China under Grant 2018GY-008, in part by the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2018JM6030 and Grant 2018JM6068, in part by the Fundamental Research Funds for the Central Universities under Grant GK201702015, in part by the Natural Sciences and Engineering Research Council of Canada and in part by the University of Alberta.

ABSTRACT Synthetic aperture imaging is a technique that mimics a camera with a large virtual convex lens with a camera array. Objects on the focal plane will be sharp and off the focal plane blurry in the synthesized image, which is the most important effect that can be achieved with synthetic aperture imaging. The property of focusing makes synthetic aperture imaging an ideal tool to handle the occlusion problem. Unfortunately, to automatically measure the focusness of a single synthetic aperture image is still a challenging problem and commonly employed pixel-based methods include using variance or using a "manual focus" interface. In this paper, a novel method is proposed to automatically determine whether or not a synthetic aperture image is in focus. Unlike conventional focus estimation methods which pick the focal plane with the minimum variance computed by the variance of corresponding pixels captured by different views in a camera array, our method automatically determines if the synthetic aperture image is focused or not from one single image of a scene without other views using a deep neural network. In particular, our method can be applied to automatically select the focal plane for synthetic aperture images. The experimental results show that the proposed method outperforms the traditional automatic focusing methods in synthetic aperture imaging as well as other focus estimation methods. In addition, our method is more than five times faster than the state-of-the-art methods. By combining with object detection or tracking algorithms, our proposed method can also be used to automatically select the focal plane that keeps the moving objects in focus. To the authors' best knowledge, it is the first time that such a method of using a deep neural network has been proposed for estimating whether or not a single synthetic aperture image is in focus.

INDEX TERMS Focusing measure, deep learning, convolutional neural network, synthetic aperture imaging.

I. INTRODUCTION

As is commonly known, typical consumer grade cameras have the ability to focus on objects located at different depths. Objects on the focal plane appear sharp while others appear blurry in images. With the reducing cost of cameras, it is now practical to use multiple cameras in a camera array for collecting different views of a scene. Synthetic aperture imaging is the technique that mimics a camera with a large virtual convex lens using a camera array. By projecting the view of each camera in a camera array onto a virtual plane located at a selected depth, a synthetic aperture image focused at the selected depth can be generated, in which some parts of the synthetic image are sharp while others blurry. The part that is sharp represents that the object is focused on the focal plane, and the part that is blurry not. Hence, finding if an object is at the focal depth corresponds to estimating whether or not the part of the image in which the object is focused.

The associate editor coordinating the review of this manuscript and approving it for publication was Bora Onat.

However, finding a good focus measure for synthetic aperture imaging is still a challenging problem. Most of the conventional focus measures used in consumer grade cameras are based on computing spatial derivatives or on using local statistics of pixel values. In the case of a camera array, multiple images of a scene from different cameras are required. As well, all of these methods usually cannot give the result based on a single synthetic aperture image but requires a sequence of such images each focusing at a different depth within a range. Because of a large amount of calculation on pixels from multiple images from different views, the conventional synthetic aperture imaging algorithms are computationally expensive.

To address the above problem in conventional focusing measures, a novel method based on a deep neural network is proposed in this paper. Deep neural networks have become very popular in recent years. They have been applied to many computer vision areas and produced impressive results. For example, Parkhi *et al.* [1] use a Convolutional Neural Network (CNN) in face recognition, which achieves an accuracy rate of 98.95%. Pathak *et al.* [2] get outstanding inpainting results in image restoration using deep learning. Krizhevsky *et al.* [3] achieve the best result in the ILSVRC2012 contest using a deep neural network. With mounting successes, deep learning or deep neural networks have become an indispensable tool in addressing problems in computer vision. Motivated by the above, the proposed method applies a deep neural network to handle the focus measure problem, and is shown to be effective in experiments.

In this paper, we propose a novel method which estimates whether or not a synthetic aperture image is focused using a CNN. Only one single synthetic aperture image is required in the proposed algorithm. The proposed VGG-16 [4] is pre-trained using ILSVRC2012 and fine tuned using a large number of synthetic aperture images. Images for fine tuning are classified into two types, namely, focused images and defocused images. When estimating whether or not an image is in focus, the features of the input image are extracted by the feature maps of the network. The probability of each class is estimated using a classifier.

The main contributions of this paper are summarized below: to the authors' best knowledge, it is the first time that such a method using a CNN has been proposed for estimating whether or not a synthetic aperture image is in focus. In addition, compared with variance-based methods which are the most commonly used in synthetic aperture imaging, our method requires only one single synthetic aperture image instead of a sequence of such images focused over a range of depths. Thus, our method takes less time compared with existing methods. Furthermore, the proposed method can be used to automatically select the optimal focal plane among the different depths that images are focused at based on the estimated probability of focus. Last but not least, an image with a small focused region can also be identified by the proposed method.

The organization of this paper is as follows. In Section II, some related works are introduced. The method we proposed is in Section III. Then we present details of the implementation in Section IV. Experimental results and performance discussion are given in Section V and VI, respectively. Finally, the paper concludes in Section VII.

II. RELATED WORKS

During the last few years, many camera array systems are built, such as the Stanford multi-camera array [5], the self-reconfigurable camera array [6] and the UCSD Eight-Camera Array [7] and so on. These systems can be classified into many categories according to their functions: high-speed videography system [8], high performance imaging system [9], image-based rendering system [10], synthetic aperture system [11]–[13], *etc.*

In the area of focus measure for images captured by a single camera, numerous works have been done which are mostly based on derivatives or local statistics of the pixel values. Most of these works measure the focus or sharpness based on variance [14], Laplacian [15], Wavelet [16] and discrete cosine transform [17]. These algorithms have been applied to solve problems such as shape from focus [15], [18], image fusion [14], [16], [19] and automatic focusing [20]. Pertuz *et al.* [18] and Hashim Mir and van Beek [21] focus measure operators into several categories according to their working principles.

Pech-Pacheco *et al.* [22] propose a method which uses the local variance of gray level to measure focus. In their opinion, a well focused image is expected to have a high variation in gray levels. The method of Pech-Pacheco *et al.* [22] computes the variance of pixels inside the region of interest. Nayar and Nakagawa [15] develop the sum-modified Laplacian operator to measure the quality of image focus. The goal of their algorithm is to increase the stability of the traditional Laplacian algorithm in shape from focus. In spite of the robustness to noise of their method, some details in images cannot be well dealt with if the size of window changes. Inspired by the fact that images with different focus levels have different marginal distributions of wavelet coefficients, Tian and Chen [16] propose an approach of sharpness measurement based on a Laplacian mixture model of wavelets. Though their result is encouraging, their method is computationally expensive. Kristan *et al.* [17] introduce a method which uses the Bayes spectral entropy of an image spectrum. The sharpness can be calculated by transforming each sub-image with a discrete cosine transform. Though the computational requirement is reduced using smaller sub-images, the region for measurement is limited by the focusing window.

In addition to applications in single cameras, focus measures have been applied to camera arrays in recent years. Yang *et al.* [23] propose an algorithm of focus measure which is based on pixel information among multiple visibility layers. However, textureless background may limit their performance. Pei *et al.* [24] propose a method of generating all-in-focus synthetic aperture images using image matting. The defocused region is replaced by focusing on background objects using energy minimization, and the focused region is sharpened using a labeling method. Their method extends the conventional method by replacing out-of-focus background with a sharp one using image matting. The shortcoming of their method is that their method may fail when the objects in the scene span a small depth range. Yang *et al.* [25] propose a novel method to solving the camera array auto-focusing problem. Moving objects through occlusion can be seen with an active camera array. However, because of the optimization and iteration in their method, the procedure is computational expensive.

Since Krizhevsky *et al.* [3] published AlexNet in 2012, more and more deep networks such as VGG [4], GoogleNet [26], ResNet [27] and DenseNet [28] have been put forward. In recent years, CNN has been applied to many areas and achieved excellent results, such as image segmentation [29]–[32], image super-resolution [33]–[37], image style transfer [38]–[41], image dehazing [42], [43], image steganography [44], etc. In particular, many image classification problems have achieved remarkable results with CNN. Parkhi *et al.* [1] use CNN for face recognition, Levi and Hassner [45] apply CNN to age and gender classification, and Narayana *et al.* [46] get outstanding results in gesture recognition with CNN, etc.

There are two differences between our work and previous related work. First, our focus measure method is based on a single synthetic aperture image while most previous methods require a sequences of such images. Second, our method is the first one to use a CNN to identify whether or not a synthetic aperture image is focused.

III. CNN-BASED FOCUS METHOD

Our focus method begins with one single synthetic aperture image. To state how the image is generated, the method of synthetic aperture imaging is introduced first. Then the deep CNN architecture used in our method follows.

A. SYNTHETIC APERTURE IMAGING

According to the plane plus parallax calibration method [11], a synthetic aperture image on an arbitrary focal plane which is parallel to the camera plane can be easily computed. Denote the reference plane which is parallel to the camera plane as π_r and the depth of the reference plane as r. Denote the N cameras in the camera array as $C_1, \cdots, C_N$ and F_i as a frame captured by camera C_i. Among these N cameras, one of them is chosen as the reference camera C_r. The homography matrix H_i warps F_i to the reference camera C_r at the reference plane π_r as shown in (1):

$$W_{i,r} = H_i \cdot F_i, \tag{1}$$

where $i = 1, \cdots, N$ and $W_{i,r}$ denotes the warped image from camera C_i to camera C_r after the homography transformation. The relative positions between cameras are represented by the displacement matrix denoted by ΔX which can be obtained from the calibration result of the camera array. Denote π_l as the target focal plane at depth l within the depth range and π_l is parallel to the reference plane π_r. When focus on the plane π_l, the ratio of the relative depths from the camera plane is denoted by $d' = (l-r)/l$. According to the method in [11], the parallax $\Delta p'$ at depth l is given by:

$$\Delta p' = \Delta X \cdot d'. \tag{2}$$

According to (2), the target plane π_l can be focused on by translating the images on the reference plane π_r using the parallax matrix $\Delta p'$. For example, assume that plane π_l at depth l is to be focused on, image $W_{i,r}$ is shifted by $\Delta p'$ using:

$$W_{i,l} = \begin{bmatrix} I & \Delta p' \\ \theta^T & 1 \end{bmatrix} W_{i,r}, \tag{3}$$

where $W_{i,l}$ denotes the shifted image focused at depth l. I is a 2×2 identity matrix. θ is a two-dimensional zero vector. Denote $W_{i,l}(q)$ as the value of pixel q. Denote $S(q)$ as the value of pixel q in the corresponding synthetic aperture image. According to (4):

$$S_l(q) = \frac{1}{N} \sum_{i=1}^{N} W_{i,l}(q), \tag{4}$$

the synthetic aperture image S_l which focuses at depth l is generated by averaging the pixel values in all warped images $W_{i,l}$.

In conclusion, to create a synthetic aperture image with the depth information, we place the calibration pattern on the reference plane π_r. Next, we project the images of different camera views onto the reference plane using homography H_i. Then, the calibration pattern is moved to different relative depths for computing the parallex. We obtain the relative camera distance ΔX with the rank-1 factorization of the matrix of parallax vectors. Based on the plane plus parallax method, with the relative camera distance ΔX, it is easy to generate parallax $\Delta p'$ using (2). Image $W_{i,l}$ is obtained by projecting $W_{i,r}$ onto the reference plane with focus at depth l. Finally, by averaging the pixel values in all warped images $W_{i,l}$ we obtain the synthetic aperture image S_l which focuses at depth l. More details can be found in [11].

The parallel parallax synthetic aperture imaging is shown in Fig. 1. P denotes a point on the target focal plane, and it has distinct imaging points p_i, p_r in cameras C_i, and C_r. Δx_i denotes the relative camera displacement between C_i and C_r, and $\Delta p'_i$ denotes the corresponding parallax.

B. DEEP CNN ARCHITECTURE

Synthetic aperture images are generated to fine tune a pre-trained VGG-16 with the fully connected and classification layers replaced. The input images are processed by hidden layers including convolution layers, activation layers, pooling layers, fully connected layers, dropout layers and the softmax layer. The network is designed to solve the problem of recognizing different classes according to features learned through mappings among layers.

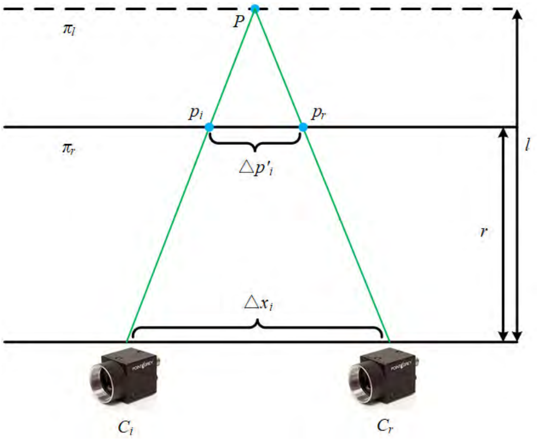

FIGURE 1. The parallel parallax synthetic aperture imaging.

Convolution layers [4], [27], [28] are used to extract features. Activation layers [3], [47] make the neural network model change from linear to nonlinear, thus being able to deal with more complex problems such as image classification. Pooling layers [48], [49] reduce the computational complexity of the network on the one hand and extract the main features on the other hand by compressing the feature maps. Fully connected layers [50] transform the two-dimensional feature maps into a one-dimensional vector, which facilitates the input of features to the final softmax layer. Droupout layers [51], [52] are to prevent overfitting in the network training process. As the final classification layer, the softmax layer [53], [54] outputs probability distribution for classification. The structure of the deep network is shown in Fig. 2. The activation layer, pooling layer, and dropout layer are not shown in Fig. 2. After an input image is processed through the different layers, the feature maps of the input image are extracted for classification. At the end of the network, the probabilities of the two classes are estimated using softmax.

Each synthetic aperture image is first resized to a resolution of 224 × 224. Next, it is processed in the convolution layer. The convolution operation is expressed as:

$$G(x,y) = \sum_{x,y}^{I}\left(\sum_{a,b}^{J} F(x+a,y+b)H(a,b)\right), \tag{5}$$

where $F(x, y)$ denotes an element of the input matrix and $G(x, y)$ an element of the output matrix. x and y denote the x-th row and the y-th column of the matrix, respectively. Similarly, a and b denote the a-th row and the b-th column of the kernel. Denote $H(a, b)$ as an element of the convolution kernel. The size of the kernel is J. I denotes the size of the input matrix. In the convolution layers, kernels are initialized by a Gaussian function. After the convolution operation of one kernel, a two-dimension feature map is generated. Hence, the output of the convolution layer is equal to the number of kernels in the layer. The detailed parameters in each convolution layer are shown in Fig.2. For example, in the convolution layer "conv1", the size of the kernel is 3 × 3 and the number of filters is 64.

In our network, there is one activation layer followed each convolution layer. The rectified linear units (ReLU) function is used as the activation function. The output of the activation layer is defined as:

$$R\left(x,y\right) = \max\left(0, G\left(x,y\right)\right), \tag{6}$$

where $R\left(x, y\right)$ denotes an element in the output matrix of the activation layer. Compared with sigmod, ReLU will speed up training with less computation.

To reduce the number of parameters, pooling layers are used in the network. In our proposed method, pooling operations are max-pooling. The output of the pooling layer is defined as:

$$U\left(x',y'\right) = \max\left(R\left(x+m,y+n\right)|m,n \in [0,\Delta I]\right), \tag{7}$$

where $U\left(x', y'\right)$ denotes an element in the output matrix of the pooling layer, and ΔI denotes the stride. In this paper, ΔI is set to 2.

Fully-connected (FC) layers are appended as "fc6", "fc7", and "fc8" as shown in Fig. 2, and each can be expressed as:

$$O\left(t\right) = \beta + \sum_{x=1}^{Q} w_{t,x} U\left(x\right), \tag{8}$$

where $O\left(t\right)$ denotes the t th element in the output matrix of the FC layer. $w_{t,x}$ denotes the weight of the xth element of the input matrix $U\left(x\right)$. β denotes the bias, and Q the number of elements in $U\left(x\right)$. The input of the first FC layer is converted

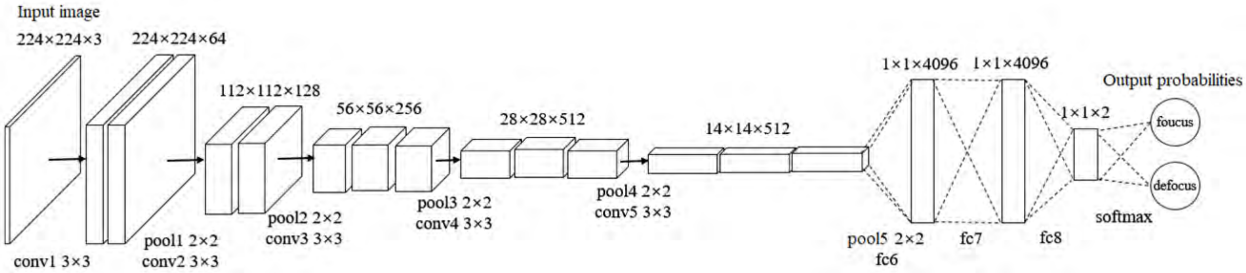

FIGURE 2. The structure of the network.

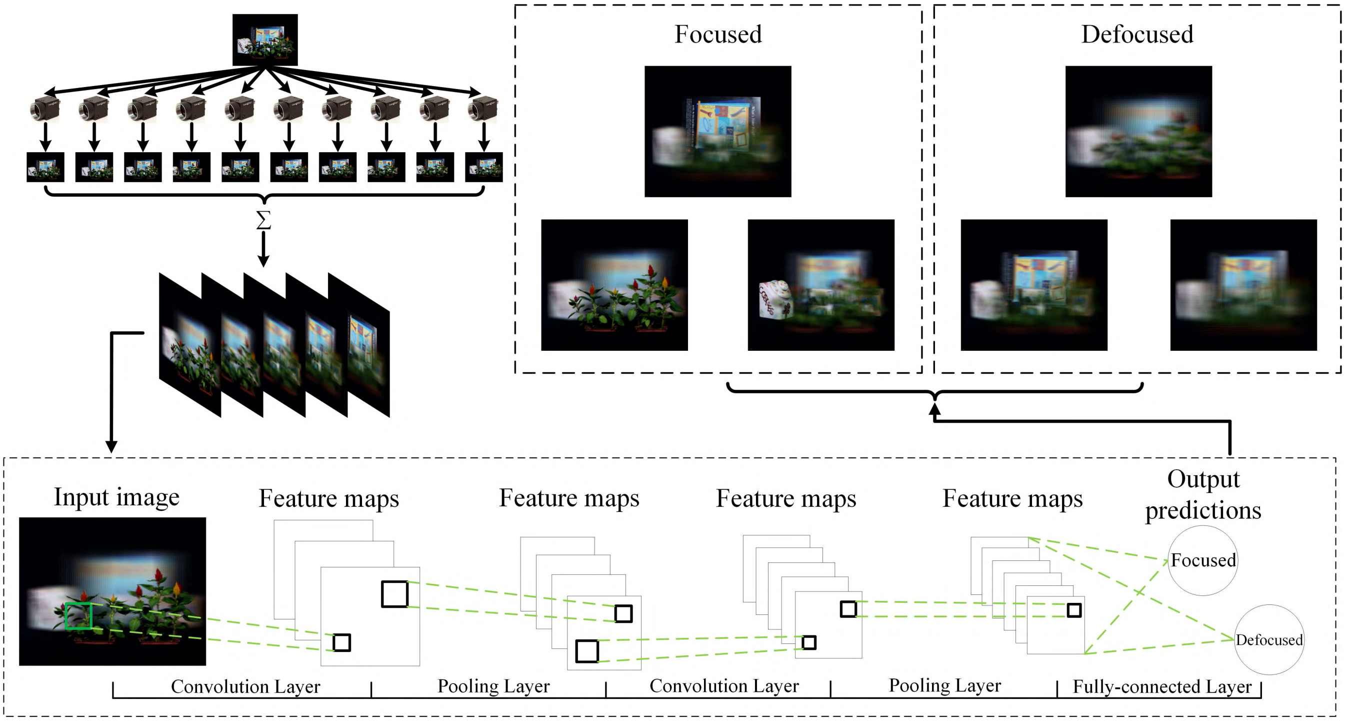

FIGURE 3. The workflow of the proposed method.

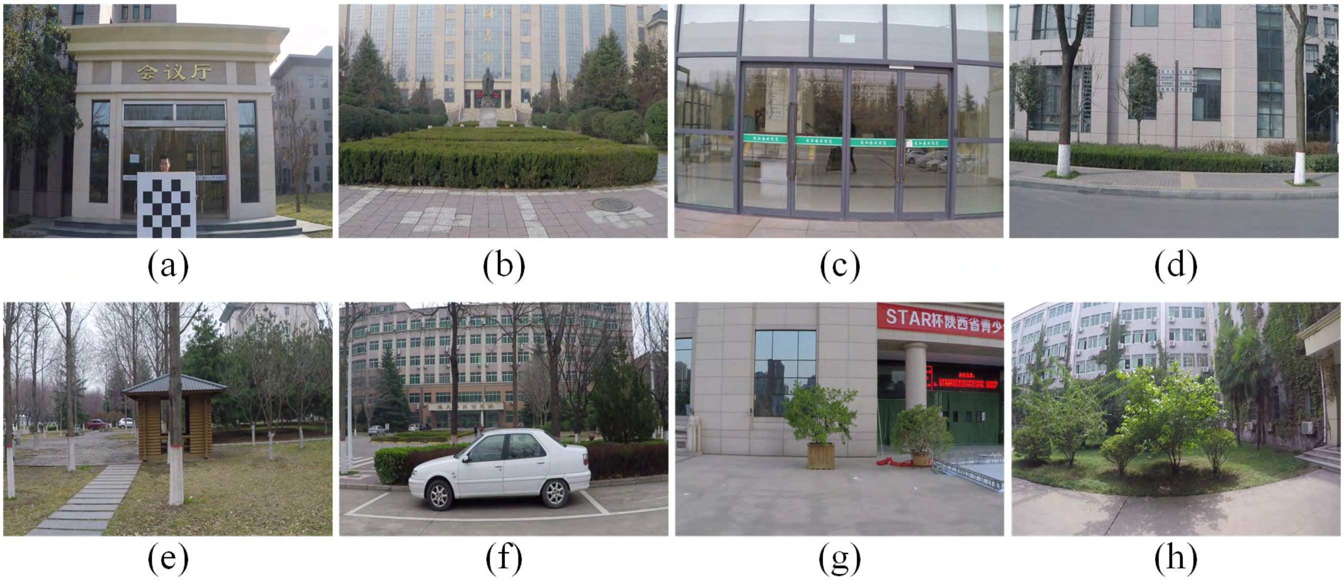

FIGURE 4. 8 scenes in our campus dataset which are used in our experiments. (c) and (f) are scenes without occlusions. The others are scenes with occluders.

by rasterizing the output of the previous layer. "fc6", "fc7", and "fc8" are fully connected. In the network, three fully-connected layers, "fc6", "fc7", and "fc8" are used. To avoid the overfitting problem in the network, dropout is used in the fully connected layers with a rate of 0.5.

To classify the output of the final fully-conntected layer into "focused" and "defocused" categories, a softmax layer is added at the end of the network. Denote p_n as the predicted probability of the k-th class in total $K = 2$ classes, and it is calculated as:

$$p_k = \frac{e^{O_k}}{\sum_{k'=1}^{K} e^{O_{k'}}}. \tag{9}$$

Then, the logistic loss is used as the network loss function, which is defined as:

$$L(w) = -\frac{1}{B}\sum_{n=1}^{B} log(p_{k,n}), \tag{10}$$

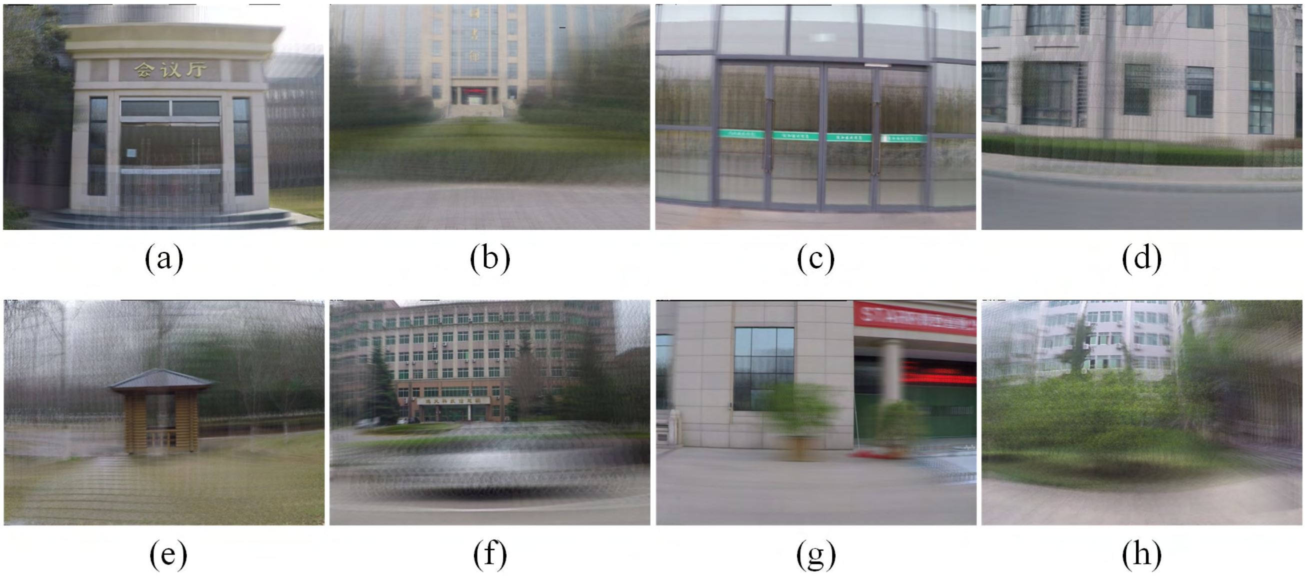

FIGURE 5. 8 focused images used in the experiments.

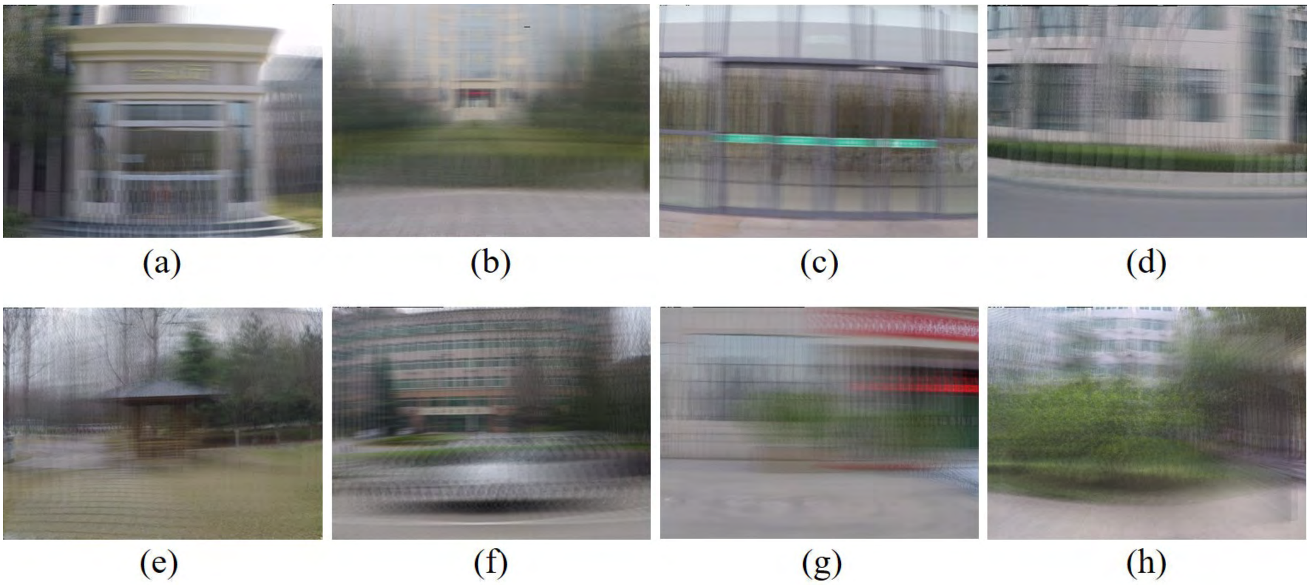

FIGURE 6. 8 defocused images used in the experiments.

where L is the loss function. B denotes the number of images in a batch in one iteration. Denote Δw as the partial derivative of the loss with respect to the weight, which is expressed by:

$$\Delta w = \frac{\partial L}{\partial w}. \tag{11}$$

The weight of a neuron is adjusted as:

$$w_{i+1} = w_i + \Delta w \cdot \alpha, \tag{12}$$

where w_i denotes the weight of the neuron in the i-th iteration. α denotes the learning rate of the network. In the proposed network, the value of α is 0.001. After thousands of iterations, the loss in the network approaches zero.

The workflow of our method is shown in Fig. 3, in which the synthetic aperture images are taken as the input of the network, processed by several hidden layers, and finally the network outputs the probabilities of focused or defocused. The pseudocode of the whole algorithm is presented in Algorithm 1.

IV. IMPLEMENTATION DETAILS

In order to get synthetic aperture imaging datasets more conveniently, a single GoPro Hero Silver 4 camera which can move horizontally on a tripod is used to simulate a camera array in the experiments. The scenes in the campus dataset are static. During the process of moving, different

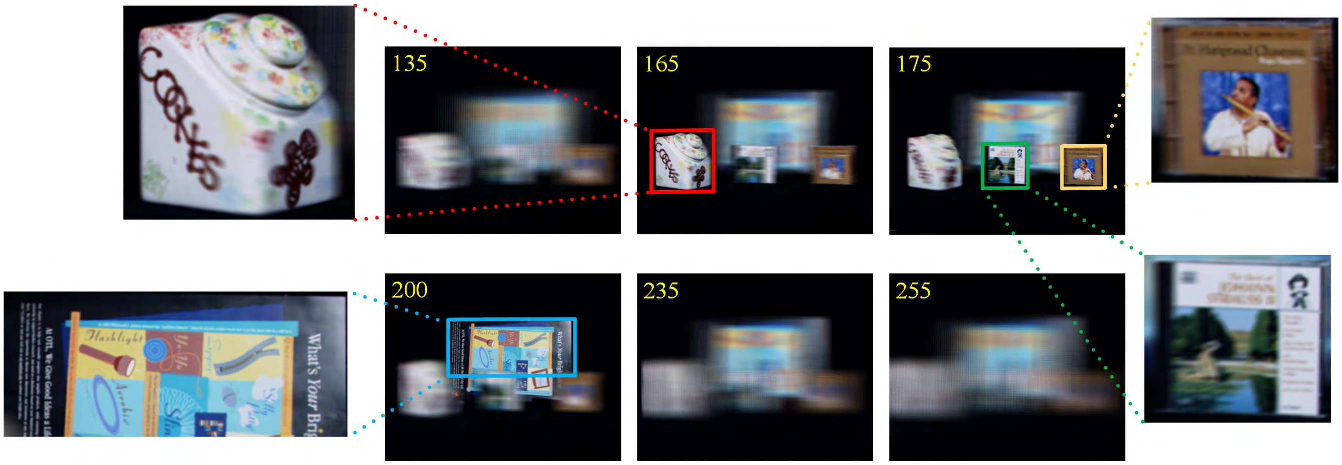

FIGURE 7. Samples of synthetic aperture images utilized in the Stanford CD-case scene. The relative focusing depth increases from top left to bottom right. The result of this scene is shown in Table 3.

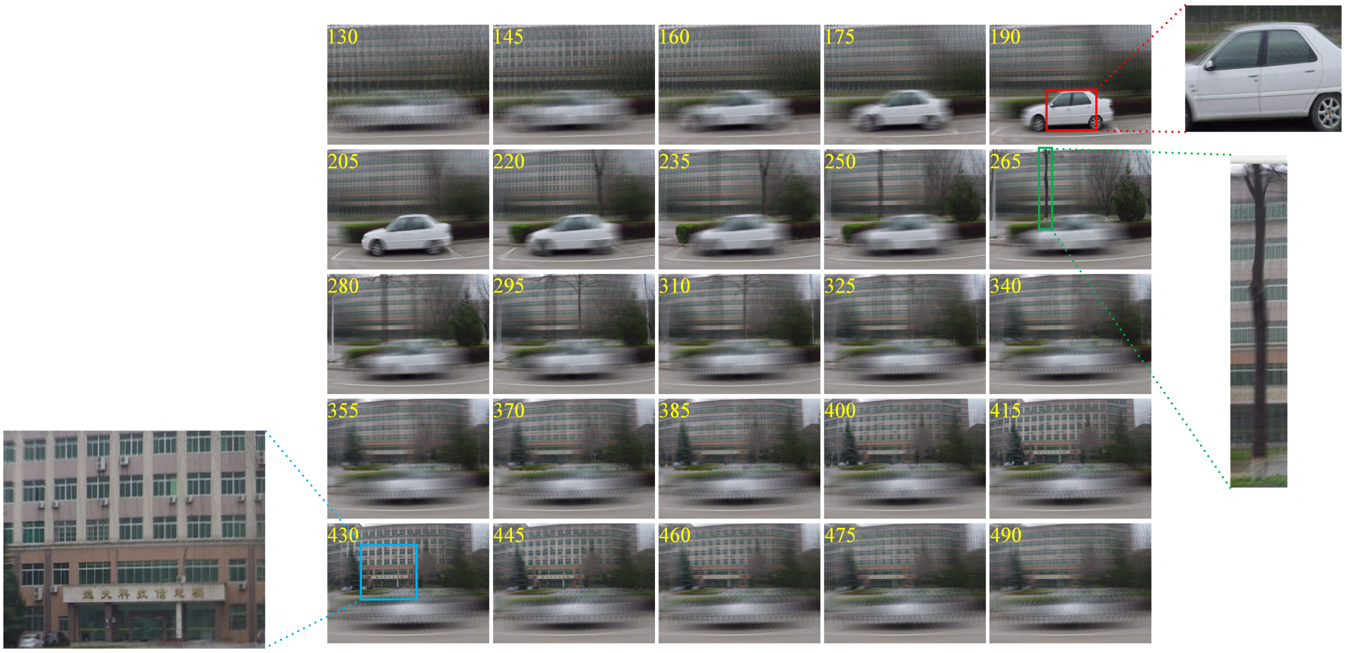

FIGURE 8. Synthetic aperture imaging results of focusing at increasing relative depths from top left to bottom right. The top left image is focused at 130 relative depth and the bottom right image is focused at 490 relative depth. Images focused at 190 depth (row 1, column 5) and 430 depth (row 5, column 1) appear sharp while others appear blurry. The curve of auto-focusing in this scene is shown in Fig. 9.

positions on the tripod can be viewed as positions of each camera in the camera array. By extracting frames in the captured video, different frames can be viewed as different positions of cameras in a camera array. The method of generating unstructured synthetic aperture images can be found in Ma *et al.*'s work [55].

After getting the data, we implement the process of synthetic aperture imaging with C++ and OpenCV 3.3.0. 44 campus scenes which include buildings, stone figures, glass walls, cars, etc. are selected as our campus dataset. In each scene, about 200 synthetic aperture images are generated at different depths for training and testing.

Caffe is used [56] as the implementation of the deep learning framework and fine-tuning is applied to the pre-trained network. VGG-16 is used in the proposed work. In the training dataset, synthetic aperture images generated from 44 scenes are classified by one of the authors into either focused or defocused. If the focused regions cover an interesting object of an image, the image is classified as focused. Otherwise, it is classified as defocused.

V. EXPERIMENTAL RESULTS

In this section, we use 2812 focused images and 288 defocused images with a resolution of 640 × 360 in 36 scenes,

Algorithm 1 Algorithm 1 of the Proposed Method

```
Generate synthetic aperture images of different scenes;
Classify generated images into two categories;
Initialize "fc6", "fc7" and "fc8" with Gaussian filter;
for each mini-batch in the training set do
    for each sample in one mini-batch do
        for each layer $i = 1$ to number of hidden layers $n$ do
            Compute the output matrix with the input matrix of
            each layer $i$ according to section III-B;
        end for
        for each layer $j = n$ to 1 do
            Compute the Loss $L_j$ according to Eq. 9 and 10;
            Update every weight $w$ in the network according to
            Eq. 11 and 12;
        end for
    end for
end for
Input testing set into the deep network;
Get probability of each category using the deep network.
if the probability of the focused class is greater than the
probability of the defocused class then
    The input image is a focused image.
else
    The input image is a defocused image.
end if
```

which are part of the 44 scenes for training, the other 8 scenes (see Fig. 4) and the Stanford CD cases dataset are used for testing.The training process runs on Ubuntu 16.04 operating system with 4.20GHZ Intel Core i7-7700K central processing unit, 32GB of memory and NVIDIA TITAN Xp graphics card. It takes 8 hours 36 minutes to train the network with 100000 iterations. Synthetic aperture images of these scenes are generated in different depths with the same interval. To evaluate the performance of determining whether or not a synthetic aperture image is focused, our method is compared with the method based on the variance of pixel values method which is widely used to distinguish the focus pixels from the background in synthetic aperture imaging, such as [9], [23], [24], [57], and [58]. Whether or not pixels in the test image are in focus is determined by the variance value. If the percentage of focused pixels is larger than a threshold of the focused region, the test image is determined as a focused one. Otherwise, the test image is determined as a defocused one. The focused and defocused images used for comparison are shown in Fig. 5 and Fig. 6, respectively, and the corresponding results are shown in Table 1 and Table 2.

The Stanford CD-cases dataset[1] consists of a scene with two CD cases and a poster and is also used for testing. The performance of estimating whether or not a single image is focused is also evaluated by the Stanford CD-cases dataset (see Fig. 7). Images focused at different depths are evaluated by the proposed method. The results are shown in Table 3.

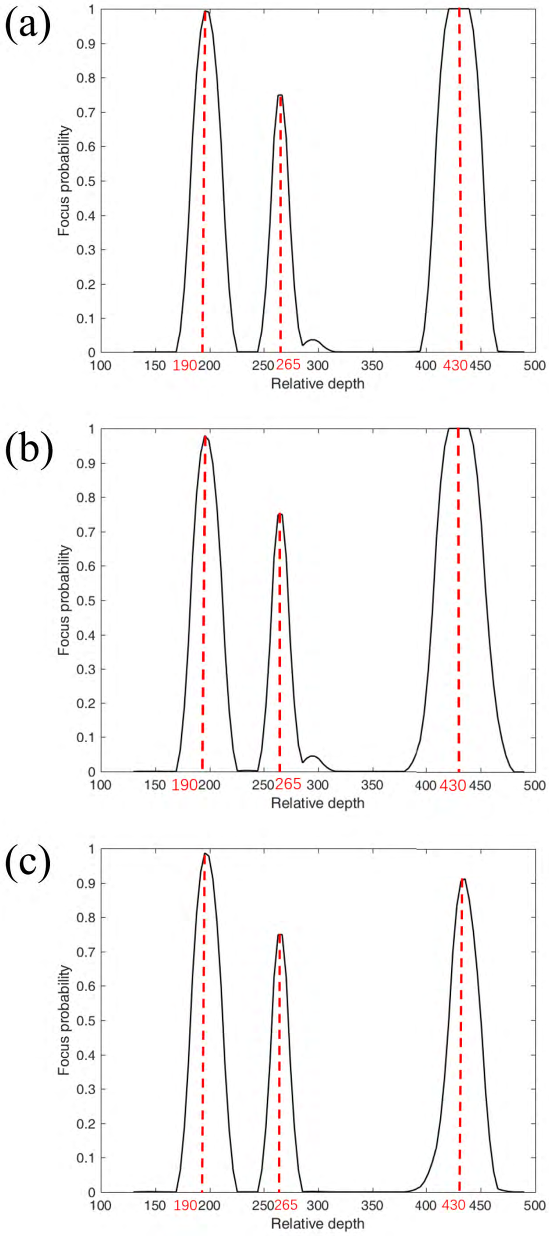

FIGURE 9. The auto-focusing result of the networks in Fig. 9. (a) The result of the avgNetwork. (b) The result of the defocusNetwork. (c) The result of the focusNetwork.

TABLE 1. The results of focused images in Fig. 5. ✓ represents that the test image is determined as a focused one. ✗ represents that the test image is determined as a defocused one.

Scene	a	b	c	d	e	f	g	h	ACC
Ours	✓	✓	✓	✓	✓	✓	✓	✓	100%
Variance	✗	✗	✗	✗	✗	✗	✓	✗	13%

The images focused at relative depths 165, 175 and 200 are well focused while images focused at depths 135, 235 and

[1] Stanford dataset download in http://lightfield.stanford.edu/lfs.html

TABLE 2. The results of defocused images Fig. 6. ✓ represents that the test image is determined as a defocused one. ✗ represents that the test image is determined as a focused one.

Scene	a	b	c	d	e	f	g	h	ACC
Ours	✓	✓	✓	✓	✓	✓	✓	✓	100%
Variance	✓	✓	✓	✓	✓	✓	✗	✓	88%

TABLE 3. The probability of the focused class of corresponding images in Fig. 7.

Depth	Probability	Depth	Probability
135	0.36%	200	100.00%
165	100.00%	235	0.12%
175	100.00%	255	0.02%

TABLE 4. Details of the number of images used for training in the three networks.

Network	Focused images	Defocused images
"avgNetwork"	1550	1550
"defocusNetwork"	288	2812
"focusNetwork"	2812	288

255 are defocused. The results show that both focused images and defocused images are estimated properly.

VI. PERFORMANCE EVALUATION AND DISCUSSION

To analyze the performance of our algorithm, the discussion is separated into three different parts. First, the influence of weights in the training set is evaluated. Then, the proposed method is compared with other focus measures for auto-focusing synthetic aperture images. Finally, the performance of our method on images with small focus regions is discussed.

A. DISCUSSING DIFFERENT WEIGHTS OF THE TRAINING SET

To evaluate the influence of the training set, the weights of focused images and defocused images are changed in three different training sets. Three CNNs which have the same structure are trained by three different training sets. In the first network, which is called the avgNetwork, the number of focused and defocused images are the same. In the second network, which is called the defocusNetwork, there are more defocused images than focused images, while in the third network, which is called the focusNetwork, there are more focused images than defocused images. Details of the number of images used for training the three networks are shown in Table 4. All three networks are trained by these three sets in the same 36 scenes with 100000 iterations. In the result of auto-focusing, three networks are evaluated by scene (f) in Fig. 4 (sequence of the synthetic aperture images are shown in Fig. 8). The three networks have different ability to select the optimal focused depth from images focused at different depths. All networks estimate relative depths around 190, 265 and 430 as the most focused images. As shown in Fig. 9, images focused at these three depths are well focused at different regions. Our method can automatically choose the optimal focal plane if the weight of training images changes.

TABLE 5. The accuracy of methods in auto-focusing. ✓ represents that the focus depth predicted by each method has the target object. ✗ represents that the focus depth predicted by each method has no target object.

Scene	Var	SML	GLV	FD	VS	RH	NV	Ours
a	✗	✓	✓	✓	✓	✗	✗	✓
b	✗	✗	✗	✓	✓	✗	✗	✗
c	✗	✗	✓	✓	✓	✗	✓	✓
d	✗	✗	✗	✗	✗	✓	✗	✓
e	✓	✓	✓	✓	✓	✓	✗	✓
f	✗	✓	✓	✓	✓	✗	✗	✓
g	✗	✓	✓	✗	✗	✗	✗	✓
h	✓	✓	✓	✓	✗	✗	✓	✓
stanford	✓	✓	✓	✓	✓	✓	✗	✓
ACC	0.33	0.67	0.78	0.78	0.67	0.33	0.22	0.89

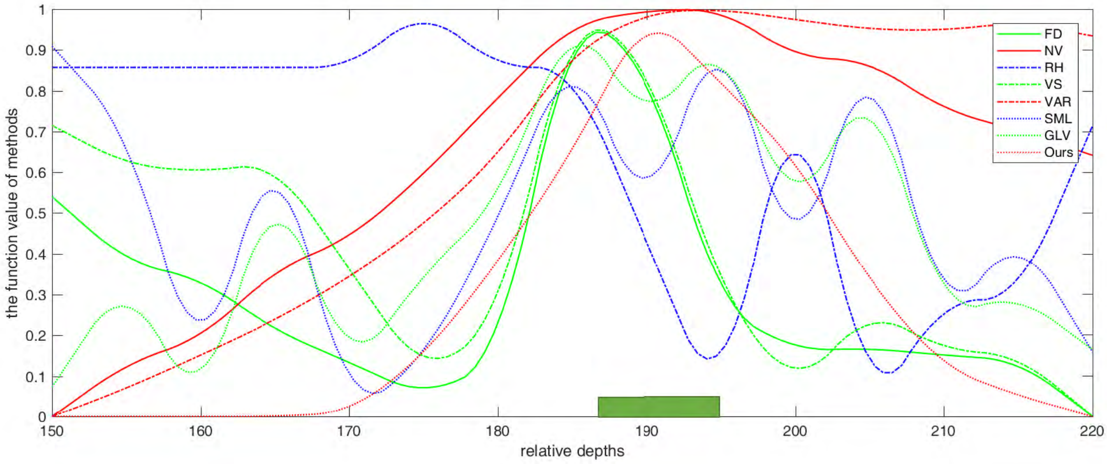

FIGURE 10. The example of comparison with scene (f) in Fig. 4. Synthetic aperture images in Fig. 8 are used in the comparison. (see text for details).

However, the performance of auto-focusing is influenced by the weights. For example, in Fig. 9 compared with (a) and (b), the result of (c) is much closer to the ground truth. Therefore, we chose the focusNetwork for final evaluation.

B. EVALUATING AUTO-FOCUSING QUALITY

The proposed method are compared with other focus measure operators in the performance of auto-focusing in Table 5, such as Sum Modified Laplacian (SML) [59], Gray-level Local Variance (GLV) [22], First-Derivative (FD) [60] which is based on first-order differentiation, Vertical Sobel (VS) [61] which is based on second-order differentiation, Range Histogram (RH) [62] which is based on image histogram and Normalized Variance (NV) [62], [63] which is based on image statistics. To explain how our method is compared with others, an example of comparison is shown in Fig. 10. Due to the different ranges of different methods, we normalize the results to a range of 0 to 1 for comparing different methods conveniently. The function value of each method represents the normalized focus measure obtained by each method. The green region above the horizontal axis denotes the optimal depth range which has the target object. In order to remove noise of small peaks, only the distance between peak and trough greater than 0.1 is considered. There are 8 scenes and the Stanford CD cases dataset used in the comparison. In each scene, around 15 synthetic aperture images focused at different depths are generated. For example, synthetic aperture images of scene (f) are generated in 150-220 depths with the interval 5. In the comparison, the range of expected focus depth is set at a relative depth of 10. Table 5 shows the comparison of the accuracy of auto-focusing using different focus measures. Fig. 11 shows autofocus images of scene (f) obtained by different focus measures. The methods based on Var, RH and NV failed to automatically focus. The methods based on SML, GLV, FD, VS all automatically focus on depth 185, which is incorrect, while our method on depth 190, which is sharp. The performance of our method is the best.

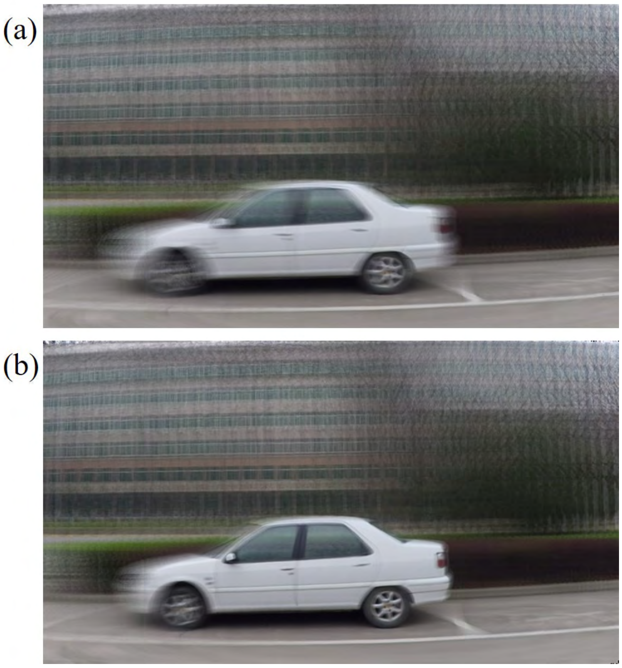

FIGURE 11. The results obtained by different focus measures. (a) is the result of the method based on SML, GLV, FD, VS. (b) is the result of our method.

The computation time of the method based on variance, SML, GLV, FD, VS, RH, NV and our method are, respectively, 960 ms, 557 ms, 465 ms, 432 ms, 960 ms, 330 ms, 419 ms and 60 ms. Thus, our method can process synthetic aperture image sequences focusing at different depths at 16.7fps, which is 5.5 times faster than the second RH method (3fps).

C. TESTING ON IMAGES WITH SMALL FOCUS REGIONS

Our method has the capacity of measuring an image which has small focus regions. The images in Fig. 12 are focused at the depth of objects with small regions. The focus probabilities of (a), (b) and (c) in Fig. 12 are, respectively, 100.00%, 100.00% and 100.00%.

TABLE 6. The computation time of the methods.

Var	SML	GLV	FD
960 ms	557 ms	465 ms	432ms

VS	RH	NV	Ours
960 ms	330 ms	419ms	60ms

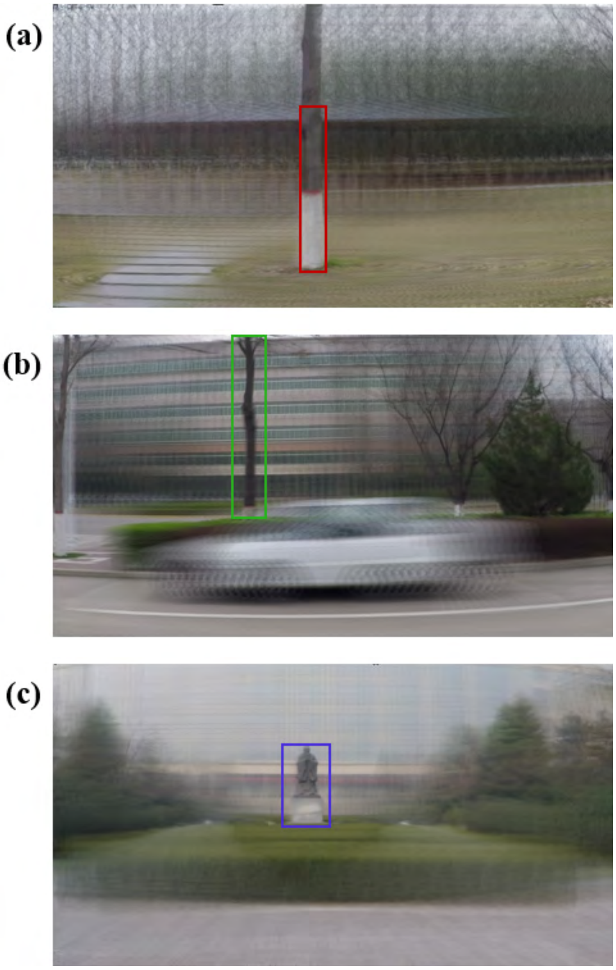

FIGURE 12. Synthetic aperture images focused on small regions. (a) and (b) Two images focused on a tree. (c) An image focused on a statue.

VII. CONCLUSION

In this paper, we propose a novel method which utilizes deep learning to estimate whether or not a synthetic aperture image is in focus. Our method uses a CNN which is a powerful method for focus measure. There are many advantages of our proposed method. First, compared to other focus measure algorithms which require image sequences, our method uses only one single synthetic aperture image as input with much less computation time. Second, our method is the first CNN-based focus measure for synthetic aperture imaging. Third, our method has the capacity of measuring an image which has small focus regions. Furthermore, our method can be used in auto-focusing. In spite of many advantages, there are still limitations in our method. According to the data-driven method, our method is sensitive to the training set. For augmentation of the training data, focal images which are near the depth of the object may decrease the performance of our method. To handle this problem, we plan to increase the number of images in the training dataset in the future. As well, developing a real-time focus measure system using deep learning for synthetic aperture images has become a possibility as demonstrated in our proposed method, which we also plan to investigate.

ACKNOWLEDGMENT

The authors would like to gratefully acknowledge the support from NVIDIA Corporation for providing them with the Titan Xp GPU used in this research.

REFERENCES

[1] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, Swansea, U.K., 2015, pp. 41.1–41.12.

[2] D. Pathak, P. Krhenbhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 2536–2544.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, 2012, pp. 1097–1105.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

[5] B. Wilburn, M. Smulski, H.-H. Kelin Lee, and M. Horowitz, "The light field video camera," in *Proc. SPIE Conf. Media Processors*, San Jose, CA, USA, 2002, pp. 29–32.

[6] C. Zhang and T. Chen, "A self-reconfigurable camera array," in *Proc. ACM Int. Conf. Exhib. Comput. Graph. Interact. Techn.*, Los Angeles, CA, USA, 2004, p. 151.

[7] N. Joshi, S. Avidan, W. Matusik, and D. J. Kriegman, "Synthetic aperture tracking: Tracking through occlusions," in *Proc. IEEE Int. Conf. Comput. Vis.*, Rio de Janeiro, Brazil, Oct. 2007, pp. 1–8.

[8] B. Wilburn, N. Joshi, V. Vaish, M. Levoy, and M. Horowitz, "High-speed videography using a dense camera array," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Washington, DC, USA, Jun. 2004, pp. 294–301.

[9] B. Wilburn, N. Joshi, V. Vaish, M. Levoy, and M. Horowitz, "High performance imaging usinglargecameraarrays," in *Proc. ACM Int. Conf. Exhib. Comput. Graph. Interact. Techn.*, Los Angeles, CA, USA, 2005, pp. 765–776.

[10] C. Lei, X. Da Chen, and Y. H. Yang, "A new multiview spacetime-consistent depth recovery framework for free viewpoint video rendering," in *Proc. IEEE Int. Conf. Comput. Vis.*, Kyoto, Japan, Sep./Oct. 2009, pp. 1570–1577.

[11] V. Vaish, B. Wilburn, N. Joshi, and M. Levoy, "Using plane + parallax for calibrating dense camera arrays," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Washington, DC, USA, Jun./Jul. 2004, pp. 2–9.

[12] V. Vaish *et al.*, "Synthetic aperture focusing using a shear-warp factorization of the viewing transform," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, San Diego, CA, USA, Sep. 2005, p. 129.

[13] Y. Qian, Y. Zheng, M. Gong, and Y. H. Yang, "Simultaneous 3D reconstruction for water surface and underwater scene," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, 2018, pp. 754–770.

[14] W. Huang and Z. Jing, "Evaluation of focus measures in multi-focus image fusion," *Pattern Recognit. Lett.*, vol. 28, no. 4, pp. 493–500, 2007.

[15] S. K. Nayar and Y. Nakagawa, "Shape from focus," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 8, pp. 824–831, Aug. 1994.

[16] J. Tian and L. Chen, "Adaptive multi-focus image fusion using a wavelet-based statistical sharpness measure," *Signal Process.*, vol. 92, no. 9, pp. 2137–2146, 2012.

[17] M. Kristan, J. Per, M. Pere, and S. Kovai, "A bayes-spectral-entropybased measure of camera focus using a discrete cosine transform," *Pattern Recognit.*, vol. 27, no. 13, pp. 1431–1439, 2006.

[18] S. Pertuz, D. Puig, and M. A. Garcia, "Analysis of focus measure operators for shape-from-focus," *Pattern Recognit.*, vol. 46, no. 5, pp. 1415–1432, 2013.

[19] I. De and B. Chanda, "Multi-focus image fusion using a morphology-based focus measure in a quad-tree structure," *Inf. Fusion*, vol. 14, no. 2, pp. 136–146, 2013.

[20] M. Subbarao and J. K. Tyan, "Selecting the optimal focus measure for autofocusing and depth-from-focus," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, no. 8, pp. 864–870, Aug. 1998.

[21] H. Mir, P. Xu, and P. van Beek, "An extensive empirical evaluation of focus measures for digital photography," *Proc. SPIE*, vol. 9023, p. 90230I, Mar. 2014.

[22] J. L. Pech-Pacheco, G. Cristóbal, J. Chamorro-Martinez, and J. Fernández-Valdivia, "Diatom autofocusing in brightfield microscopy: A comparative study," in *Proc. Int. Conf. Pattern Recognit.*, Barcelona, Spain, 2000, pp. 314–317.

[23] T. Yang *et al.*, "All-in-focus synthetic aperture imaging," in *Proc. Eur. Conf. Comput. Vis.*, Zurich, Switzerland, 2014, pp. 1–15.

[24] Z. Pei, X. Chen, and Y.-H. Yang, "All-in-focus synthetic aperture imaging using image matting," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 2, pp. 288–301, Feb. 2018.

[25] T. Yang *et al.*, "Simultaneous active camera array focus plane estimation and occluded moving object imaging," *Image Vis. Comput.*, vol. 32, no. 8, pp. 510–521, 2014.

[26] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 1–9.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[28] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," in *Proc. IEEE CVPR*, vol. 1, Jun. 2017, p. 3, no. 2.

[29] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 3431–3440.

[30] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[31] Z. Li, Q. Chen, and V. Koltun, "Interactive image segmentation with latent diversity," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 577–585.

[32] Z. Liu, C. Cao, S. Ding, Z. Liu, T. Han, and S. Liu, "Towards clinical diagnosis: Automated stroke lesion segmentation on multi-spectral MR image using convolutional neural network," *IEEE Access*, vol. 6, pp. 57006–57016, 2018.

[33] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 1646–1654.

[34] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, vol. 1, no. 2, Jul. 2017, p. 5.

[35] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 2472–2481.

[36] L. Zhao, Q. Sun, and Z. Zhang, "Single image super-resolution based on deep learning features and dictionary model," *IEEE Access*, vol. 5, pp. 17126–17135, 2017.

[37] Y. Liu, X. Xu, J. Xu, and Z. Jiang, "Image super-resolution reconstruction based on disparity map and CNN," *IEEE Access*, vol. 6, pp. 53489–53498, 2018.

[38] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 2414–2423.

[39] X. Huang and S. J. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. Int. Conf. Comput. Vis.*, Venice, Italy, 2017, pp. 1510–1519.

[40] A. Selim, M. Elgharib, and L. Doyle, "Painting style transfer for head portraits using convolutional neural networks," *ACM Trans. Graph.*, vol. 35, no. 4, p. 129, 2016.

[41] C. Li and M. Wand, "Combining markov random fields and convolutional neural networks for image synthesis," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 2479–2486.

[42] J. Li, G. Li, and H. Fan, "Image dehazing using residual-based deep CNN," *IEEE Access*, vol. 6, pp. 26831–26842, 2018.

[43] C. Li, J. G. Guo, F. Porikli, H. Fu, and Y. Pang, "A cascaded convolutional neural network for single image dehazing," *IEEE Access*, vol. 6, pp. 24877–24887, 2018.

[44] D. Hu, L. Wang, W. Jiang, S. Zheng, and B. Li, "A novel image steganography method via deep convolutional generative adversarial networks," *IEEE Access*, vol. 6, pp. 38303–38314, 2018.

[45] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 34–42.

[46] P. Narayana, J. R. Beveridge, and B. A. Draper, "Gesture recognition: Focus on the hands," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 5235–5244.

[47] B. Xu, N. Wang, T. Chen, and M. Li. (2015). "Empirical evaluation of rectified activations in convolutional network." [Online]. Available: https://arxiv.org/abs/1505.00853

[48] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. Audio, Speech Language Process.*, vol. 22, no. 10, pp. 1533–1545, Oct. 2015.

[49] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 649–657.

[50] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 1891–1898.

[51] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[52] G. E. Hinton, N. Srivastave, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. (2012). "Improving neural networks by preventing co-adaptation of feature detectors." [Online]. Available: https://arxiv.org/abs/1207.0580

[53] Y. Tang. (2013). "Deep learning using linear support vector machines." [Online]. Available: https://arxiv.org/abs/1306.0239

[54] Z. Yin, B. Wan, F. Yuan, X. Xia, and J. Shi, "A deep normalization and convolutional neural network for image smoke detection," *IEEE Access*, vol. 5, pp. 18429–18438, 2017.

[55] W. Ma, T. Yang, Y. Zhang, and X. Tong, "Unstructured synthetic aperture photograph based occluded object imaging," in *Proc. Int. Conf. Image Graph.*, Qingdao, China, 2013, pp. 34–39.

[56] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia*, Orlando, FL, USA, 2014, pp. 675–678.

[57] Z. Pei, Y. Zhang, T. Yang, X. Zhang, and Y.-H. Yang, "A novel multiobject detection method in complex scene using synthetic aperture imaging," *Pattern Recognit.*, vol. 45, no. 4, pp. 1637–1658, 2012.

[58] Z. Pei, Y. Zhang, X. Chen, and Y.-H. Yang, "Synthetic aperture imaging using pixel labeling via energy minimization," *Pattern Recognit.*, vol. 46, no. 1, pp. 174–187, 2013.

[59] A. Thelen, S. Frey, S. Hirsch, and P. Hering, "Improvements in shape-from-focus for holographic reconstructions with regard to focus operators, neighborhood-size, and height value interpolation," *IEEE Trans. Image Process.*, vol. 18, no. 1, pp. 151–157, Jan. 2009.

[60] J. F. Brenner, B. S. Dew, J. B. Horton, T. King, P. W. Neurath, and W. D. Selles, "An automated microscope for cytologic research," *J. Histochem. Cytochem.*, vol. 24, no. 1, pp. 100–111, 1971.

[61] E. Krotkov, "Focusing," *Int. J. Comput. Vis.*, vol. 1, no. 3, pp. 223–237, 1988.

[62] F. C. A. Groen, I. T. Young, and G. Ligthart, "A comparison of different focus functions for use in autofocus algorithms," *Cytometry*, vol. 6, no. 2, pp. 81–91, 1985.

[63] S. Yousefi, M. Rahman, N. Kehtarnavaz, and M. Gamadia, "A new autofocus sharpness function for digital and smart-phone cameras," in *Proc. IEEE Int. Conf. Consum. Electron.*, Las Vegas, NV, USA, Jun. 2011, pp. 475–476.

ZHAO PEI received the B.E., M.S., and Ph.D. degrees from Northwestern Polytechnical University, Xi'an, China, in 2005, 2008, and 2013, respectively. He was a joint Ph.D. Student with the Department of Computing Science, University of Alberta, Edmonton, Canada, from 2010 to 2011. He is currently an Associate Professor with the School of Computer Science, Shaanxi Normal University, Xi'an, and he is also a Visiting Professor with the Department of Computing Science, University of Alberta. His research interests include camera array synthetic aperture imaging, object detection and tracking, and human body motion analysis.

LI HUANG is currently with the School of Computer Science, Shaanxi Normal University. Her research interests include computer vision and pattern recognition.

YANNING ZHANG is currently with the School of Computer Science, Northwestern Polytechnical University. She has published over 200 papers in international journals and conferences. Her research work focuses on signal and image processing, computer vision, and pattern recognition.

MIAO MA is currently with the School of Computer Science, Shaanxi Normal University. Her research interests include image processing and machine learning.

YALI PENG is currently with the School of Computer Science, Shaanxi Normal University. Her research interests include computer vision and machine learning.

YEE-HONG YANG received the B.Sc. degree (Hons.) from The University of Hong Kong, the M.Sc. degree from Simon Fraser University, and the M.S.E.E. and Ph.D. degrees from the University of Pittsburgh. He was a Faculty Member with the Department of Computer Science, University of Saskatchewan, from 1983 to 2001, and was the Graduate Chair, from 1999 to 2001. In addition to the department-level committees, he was also with many college and university level committees. From 2003 to 2005, he was the Associate Chair (Graduate Studies) with the Department of Computing Science, University of Alberta, where he has been a Professor, since 2001. He has authored or co-authored over 140 papers in international journals and conference proceedings in the areas of computer vision and graphics. His research interest includes a wide range of topics from computer graphics to computer vision. He has served on the program committees of many national and international conferences. In 2007, he was invited to serve on the Expert Review Panel to evaluate computer science research in Finland. He serves on the Editorial Board of the journal *Pattern Recognition*. He served as a Reviewer for numerous international journals, conferences, and funding agencies.