Composite deep learning framework for absolute 3D shape measurement based on single fringe phase retrieval and speckle correlation

Journal of Physics: Photonics

PAPER

Composite deep learning framework for absolute 3D shape measurement based on single fringe phase retrieval and speckle correlation

Wei Yin[1,2,3], Jinxin Zhong[1,2,3], Shijie Feng[1,2,3], Tianyang Tao[1,2,3], Jing Han[1,2], Lei Huang[4], Qian Chen[1,2] and Chao Zuo[1,2,3]

[1] School of Electronic and Optical Engineering, Nanjing University of Science and Technology, No. 200 Xiaolingwei Street, Nanjing, Jiangsu Province 210094, People's Republic of China
[2] Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense, Nanjing University of Science and Technology, Nanjing, Jiangsu Province 210094, People's Republic of China
[3] Smart Computational Imaging (SCI) Laboratory, Nanjing University of Science and Technology, Nanjing, Jiangsu Province 210094, People's Republic of China
[4] Brookhaven National Laboratory, NSLS II 50 Rutherford Drive, Upton, New York 11973-5000, United States of America

E-mail: zuochao@njust.edu.cn **and** surpasszuo@163.com

Keywords: phase retrieval, stereo matching, deep learning, fringe projection profilometry, speckle matching,

OPEN ACCESS

RECEIVED
26 March 2020

REVISED
12 June 2020

ACCEPTED FOR PUBLICATION
30 September 2020

PUBLISHED
22 October 2020

Abstract

Fourier transform profilometry (FTP) is a classic three-dimensional (3D) shape measurement technique that can retrieve the wrapped phase from a single fringe pattern. However, suffering from the spectral leakage and overlapping problems, it generally yields a coarse phase map with low spatial resolution and precision. Recently, deep learning has been introduced to the field of Fringe projection profilometry (FPP), revealing promising results in fringe analysis, phase unwrapping, depth constraint and system calibration. However, for absolute shape measurement of general objects, the inherent depth ambiguity problem of a single fringe is still insurmountable. In this work, we propose a composite deep learning framework for absolute 3D shape measurement based on single fringe phase retrieval and speckle correlation. Our method combines the advantages of FPP techniques for high-resolution phase retrieval and speckle correlation approaches for robust unambiguous depth measurement. The proposed deep learning framework comprises two paths: one is a U-net-structured network, which is used to extract the wrapped phase maps from a single fringe pattern with high accuracy (but with depth ambiguities). The other stereo matching network produces the initial absolute (but with low resolution) disparity map from an additional speckle pattern. The initial disparity map is refined by exploiting the wrapped phase maps as an additional constraint and finally, a high-accuracy high-resolution disparity map for absolute 3D measurement can be obtained. Experimental results demonstrated that the proposed deep-learning-based method could realize high-precision absolute 3D measurement with an accuracy of 50 μm for measuring objects with complex surfaces.

1. Introduction

Fringe projection profilometry (FPP), as an optical three-dimensional (3D) shape measurement method, is one of the most key and new-fashioned techniques in optical metrology, which can be widely applied in various fields such as machine vision, automatic optical inspection, visual inspection, industrial quality control, life science and other scientific research fields. Due to the nature of non-contact and accuracy measurement, many FPP-based methods have been developed for measuring static objects to obtain high-precision 3D reconstruction results with full-resolution and full-field [1–6]. Recently, with the rapid development of high-speed cameras and high-speed DMD-based digital light processing (DLP) projection technology, it makes some FPP-based methods have the potential to achieve high-speed and high-accuracy 3D measurements of dynamic scenes [7–10].

In general, there are three main processing steps in FPP: phase extraction, phase unwrapping and phase-to-height mapping. During phase recovery, the use of the sinusoidal fringe pattern is more prevalent to retrieval the wrapped phase using Fourier transform methods in frequency domain [11] or phase-shifting methods in time domain [12]. Phase-shifting profilometry (PSP) is capable of full-resolution and high-accuracy phase measurement, but it is not suitable for dynamic measurement due to the requirement of at least three fringes [12]. Numerous dynamic 3D measurement systems have been developed based on Fourier transform profilometry (FTP), which has the advantage of providing the phase map utilizing only a single high-frequency fringe pattern [11, 13]. However, suffering from the spectrum overlapping problem, these methods generally yield a coarse wrapped phase with low quality, which limits its measurement precision for dynamic 3D acquisition. Recently, Feng *et al* [14, 15] presented a deep-learning-based fringe analysis method that employs deep neural networks to recovery the high-precision phase information. Nevertheless, this method can only be used to almost perfectly extract the wrapped phase from a single fringe pattern with extremely high frequency ($\geq$100), which brings about more phase ambiguities and makes it difficult to successfully implement phase unwrapping for absolute 3D measurement. And these phase unwrapping methods can be grouped into three main classes in aspects of operating domain: spatial phase unwrapping [16, 17], temporal phase unwrapping [18–22], stereo phase unwrapping [23–25]. Spatial phase unwrapping is highly suited for dynamic 3D acquisition and can provide the relatively absolute phase map using only a single wrapped phase [16]. However, the continuity of the phase is an essential prerequisite for the successful implementation of spatial phase unwrapping, making it impossible for measuring discontinuous surfaces or abrupt depth with step heights greater than 2π. In order to solve the problem above, temporal phase unwrapping methods are proposed to realize absolute phase unwrapping with the aid of additional wrapped phase maps with different frequency [18].

Recently, in addition to the two methods above, stereo phase unwrapping methods in 3D measurements are more popular due to potentially addressing the respective shortcomings of spatial phase unwrapping and temporal phase unwrapping. Many stereo phase unwrapping algorithms are proposed by introducing another camera into the existing FPP system, which can retrieve the absolute phase using the single high-frequency wrapped phase maps obtained from different perspectives without projecting any additional patterns. Zhong *et al* [23] introduced the trifocal tensor constraint into PSP to search the corresponding point of each pixel independently only in the wrapped phase map, performing fast 3D measurement of arbitrary shape objects using only three fringe patterns. However, in order to ensure the stability of stereo phase unwrapping, the period of fringe is generally around 20, which limits the accuracy of 3D measurement. It can be concluded that using more cameras can provide more geometric constraints to completely eliminate the phase ambiguities in the multi-view system. Following this idea, Tao *et al* [26] proposed a position-optimized quad-camera system to guarantee the reliability of phase unwrapping for the 48-period wrapped phases. However, the valid measured 3D volumes, defined as the common regions of the cameras and projector, were significantly reduced. On the other hand, by additionally projecting auxiliary patterns or embedding the auxiliary signal into original fringe patterns, the extra auxiliary information is obtained to assist the absolute phase recovery. For the wrapped phases with N-period fringes, the fringe order of each valid pixel exists N possibilities. Taking each possible order into consideration, the corresponding depth value can be calculated using calibration parameters between the primary camera and the projector. It is easy to find that some possible depth values are beyond the pre-defined depth range based on depth constraint and the corresponding orders can be excluded from the candidates. For the remaining ones, the block matching operation based on auxiliary information will be implemented to further remove the wrong candidates. Tao *et al* [27] proposed a real-time 3D shape measurement method by embedding the triangular wave in the original phase-shifting fringes under the guidance of the number theory to ensure stereo phase unwrapping for each point. Yin *et al* [28] developed an optimized composite fringe patterns by selecting speckles as auxiliary information. Besides, the simple and effective evaluation criterion for the designed speckle pattern was introduced to improve the matching accuracy significantly. Since block matching is only performed to distinguish a few periodic candidate points, these methods enable efficient pixel-wise stereo phase unwrapping, but it is still inevitable, especially for exist fringe order errors around dark regions and object edges where the fringe quality is low.

In order to overcome the problems mentioned above, Lohry *et al* [29] presented a new stereo-phase-based absolute 3D shape measurement that requires neither phase unwrapping nor projector calibration. This method can be divided into two steps: obtain a coarse disparity map from the quality map and refine the disparity map using wrapped phases. Song *et al* [30] presented a simple, fast 3D shape measurement method using FTP, which introduces the binocular stereo vision and exploits two image pairs (i.e. original image pairs and fringe image pairs) to restructure the 3D shape of the tested object. A coarse disparity map from left and right original images is firstly calculated using the Efficient Large-Scale Stereo Matching (ELAS) algorithm and can be adopted as a disparity constraint to acquire the refinement disparity

with subpixel precision using the wrapped phase without phase unwrapping. Obviously, phase unwrapping is no longer a necessary process in the stereo phase unwrapping method but stereo matching. Similarly, stereo matching is also one of the classic tasks in computer vision. Conventional stereo matching methods, which are based on local correlation methods, semi-global matching (SGM) methods, or global matching methods, are usually implemented to build the global correspondence of stereo images and obtain the dense disparity map. For FPP, a large number of stereo matching algorithms have been proposed to enhance the accuracy and computational efficiency of stereo matching and acquire the disparity map with subpixel precision by exploiting various constraints information, including phase constraint, geometric constraint and depth constraint [31–33]. However, the universality and precision of these methods are still not enough to meet high-precision and high-efficient 3D measurement applications.

Deep learning is a powerful machine learning method that has achieved great success in numerous fields such as machine vision, image processing, natural language processing and speech recognition [34]. As a most popular deep neural network for computer vision, convolutional neural networks (CNNs) have shown its excellent performance with high robustness for four basic application tasks in computer vision, including image classification, object detection, semantic segmentation and instance segmentation. Recently, CNNs have been adopted to many 3D vision applications such as stereo vision [35–40], optical flow estimation [41, 42], 3D object detection [43, 44] and 3D tracking [45]. In the main task of stereo vision, the Siamese network is the most frequently used neural network framework, which takes left and right image patches to feed a weight-shared convolutional subnetwork and adopt some fully connected layers to predict final matching costs. Based on the Siamese network framework, LeCun *et al* [35] have carefully investigated different CNN based architectures and proposed two networks for different purposes (for speed and accuracy) to perform stereo matching. In addition, training the network is implemented to maximize the accuracy of matching in a supervised learning way by constructing a binary classification task, where the stereo matching dataset is only divided into similar pairs of patches (positive samples) and dissimilar pairs of patches (negative samples). However, for the stereo vision systems with a wide baseline, there are a large number of candidate points when implementing matching operation along the epipolar line, resulting in a serious imbalance between the positive and negative samples in the binary classification dataset. To address this problem, Luo *et al* [36] exploited left and right image patches with different sizes to train the similar CNNs, converting the binary classification problem into a multi-classification task. This method can directly predict the correlation value (that is the initial matching cost) of all candidates to be matched, making high-efficient stereo matching possible. Although these learning-based stereo matching methods have improved the matching accuracy significantly compared with the traditional method, the raw output of the CNN is not enough to produce accurate disparity maps and matching errors are still inevitable, especially around low-texture regions and object edges. Therefore, most methods set the output of the network as the initial disparity map and enhance this result by implementing a series of post-processing procedures including cross-based cost aggregation, SGM, left-right consistency check, subpixel disparity refinement, median filtering and bilateral filtering. In recent years, some new stereo matching networks have been proposed, which directly or implicitly integrate cost aggregation and SGM into the existing networks and outperformed the state-of-the-art methods on stereo vision datasets [37, 38, 40].

In this work, we propose a composite deep learning framework for absolute 3D shape measurement based on single fringe phase retrieval and speckle correlation. First, an improved U-net network, which is used for phase retrieval from one 57-period fringe image, is proposed to output the wrapped phases without background information. Since the trained network is only focused on recovering the phase information in the foreground, it enables our method to generate a more accurate phase. Besides, we adopt deep neural networks to beat the stereo matching problem for FPP, which substantially promotes the matching reliability and efficiency compared with the state-of-the-art stereo matching methods. Specifically, an additional speckle pattern is projected and input the new stereo matching network to obtain a relatively accurate matching result as the initial disparity map. The wrapped phases predicted by the phase retrieval network as the enhanced phase constraint can be exploited to refine the initial disparity map to dense and accuracy disparity maps for absolute 3D measurement without phase unwrapping. Experimental results demonstrated the proposed deep-learning-based method using only two projection patterns can realize high-precision absolute 3D measurement with an accuracy of 50 μm for objects with complex surfaces.

2. Principle

2.1. Phase retrieval through the U-net network

In this section, an improved U-net network, which is used for phase retrieval from one high-frequency fringe image, is proposed to output the wrapped phases without background information. Since the trained network is only focused on recovering the phase information in the foreground, it enables our method to

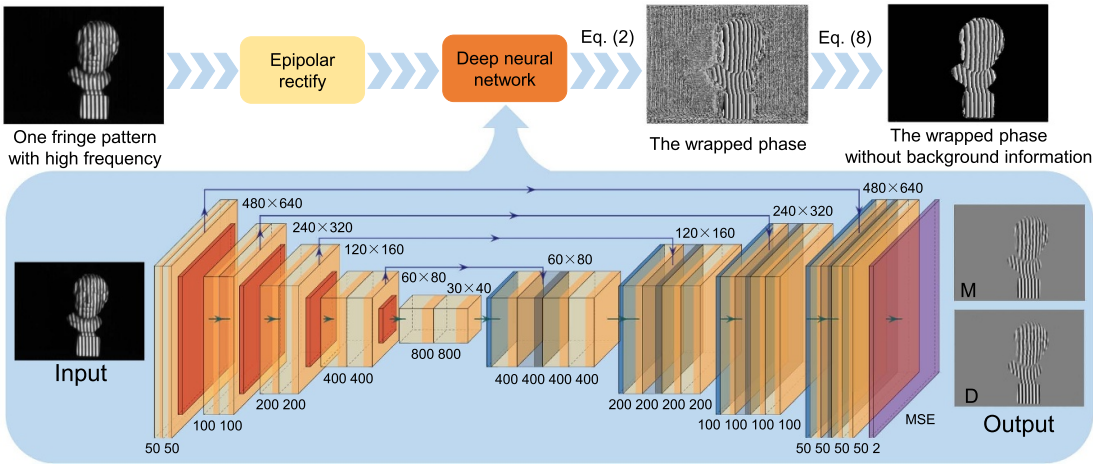

Figure 1. The diagram of the proposed phase retrieval method. The whole framework is composed of epipolar rectify, deep neural network, phase calculation and background removal. Deep neural network, as a modified U-net network, is used to predict the numerator $M(x, y)$ and the denominator $D(x, y)$ from one fringe pattern with high frequency. Then, the wrapped phase without background information is obtained using equations (2) and (8).

generate a more accurate phase. Specifically, the specific diagram of the proposed phase retrieval network is shown as in figure 1. As the input of the phase retrieval network, the fringe pattern with high frequency, which was captured by the camera, can be described as

$$I^c(x,y) = A(x,y) + B(x,y)\cos\phi(x,y), \tag{1}$$

where $I^c(x,y)$ represent the intensity of captured images, $A(x,y)$, $B(x,y)$ and $\phi(x,y)$ are the average intensity, the intensity modulation and the phase distribution of the measured object. Whatever FTP or PSP, the wrapped phase map $\phi(x,y)$ can be obtained from the uniform equation:

$$\phi(x,y) = \tan^{-1}\frac{M(x,y)}{D(x,y)} = \tan^{-1}\frac{\rho B(x,y)\sin\phi(x,y)}{\rho B(x,y)\cos\phi(x,y)}, \tag{2}$$

where $M(x,y)$ and $D(x,y)$ denote the numerator and the denominator of the arctan function, respectively. ρ is a constant that depends on the phase demodulation algorithm, e.g. $\rho = 0.5$ for FTP and $\rho = N/2$ for N-step PSP. Although the purpose of building the network is to achieve phase retrieval and obtain the wrapped phase, there is no need to directly set the wrapped phase as the network's label because the arctangent curve is abrupt and difficult to learn. According to equation (2), $M(x,y)$ and $D(x,y)$ will be set as the output data of the network. It is easy to understand that the complexity of the network will be greatly reduced so that the loss of the network will converge faster and more stable and the prediction accuracy of the network is effectively improved.

In our work, U-net is adopted to complete the prediction of the wrapped phase. The input and output of U-net are all images and there is no dense layer in the network, which can combine the low-level and high-level information at the same time [46]. The low-level information is helpful to improve the accuracy and the high-level information is used to extract complex and abstract features, which can achieve highly accurate results predicted by the U-net network. For this phase retrieval network, the ground truth can be acquired using N-step phase-shifting fringe patterns, like

$$I_n^p(x,y) = 0.5 + 0.5\cos(2\pi fx - 2\pi n/N), \tag{3}$$

where $I_n^p(x,y)(n = 0, 1, 2, \ldots, N-1)$ represents fringe patterns projected by the projector, f is the frequency of fringe patterns. Then the fringe images captured by the camera can be described as

$$I_n^c(x,y) = A^c(x,y) + B^c(x,y)\cos(\phi^c(x,y) - 2\pi n/N), \tag{4}$$

where $I_n^c(x,y)$, $A^c(x,y)$, $B^c(x,y)$ and $\phi^c(x,y)$ is same to the meanings in equation (1). According to the least-squares algorithm, the wrapped phase $\phi^c(x,y)$, $B^c(x,y)$ and $Mask_v^c(x,y)$ can be obtained:

$$\phi^c(x,y) = \tan^{-1}\frac{\sum_{n=0}^{N-1} I_n^c(x,y)\sin(2\pi n/N)}{\sum_{n=0}^{N-1} I_n^c(x,y)\cos(2\pi n/N)}, \tag{5}$$

$$B^c(x,y) = \frac{2}{N}\sqrt{\left[\sum_{n=0}^{N-1} I_n^c(x,y)\sin(2\pi n/N)\right]^2 + \left[\sum_{n=0}^{N-1} I_n^c(x,y)\cos(2\pi n/N)\right]^2}, \tag{6}$$

$$Mask_v^c(x,y) = B^c(x,y) > Thr1, \tag{7}$$

where $Thr1$ is the preset threshold for the tested object, $Mask_v^c(x,y)$ can be used to identify the valid points. The threshold $Thr1$ should be changed for object surfaces with different reflectivity, theoretically. In most cases, $Thr1 = 0.01$ is acceptable for various objects in our measurement. In our method, $Mask_v^c(x,y)$ is exploited to preprocess the ground truth for enhancing the learning ability of the network to the valid information of the measured scenes. Once the ground truth without background information is set as the output data of the network, it can be found that the prediction results output directly by the U-net network has valid values only in the foreground and negligible values in the background as in figure 1. And the prediction results without background information are obtained by following $Mask_{MD}(x,y)$ operation:

$$Mask_{MD}(x,y) = abs(M(x,y)) < 0.002 \;\&\; abs(D(x,y)) < 0.002, \tag{8}$$

It is shown from figure 1 that our proposed phase retrieval method through the U-net network can successfully extract high-precision wrapped phases without background information used as the enhanced phase constraints in the next subsection.

2.2. The deep neural network for stereo matching

In this section, we will adopt deep neural networks to beat the stereo matching problem for FPP, which substantially promotes the matching reliability and efficiency compared with the state-of-the-art stereo matching methods. There is generally a four-step pipeline for stereo matching, including matching cost calculation, cost aggregation, disparity computation and disparity refinement [47]. Traditional stereo matching methods perform all four steps using non-learning techniques. Existing learning-based stereo matching methods attempt to exploit deep learning to solve one or multiple of the four steps to obtain better matching results. For the dataset built in this work, the label of the sample data only has valid values in the foreground shown as in figures 1 and 2. Due to the characteristics of this dataset, it is difficult to simply integrate the end-to-end networks [38, 40] into our work to directly obtain the final disparity map, but implement matching cost calculation using the networks [35, 36]. Specifically, an additional speckle pattern is projected and input the new stereo matching network to obtain a relatively accurate matching result as the initial disparity map. The CNNs are adopted as the basic skeleton of our deep neural network which can provide better matching results based on one optimally designed speckle pattern and the specific diagram of the proposed stereo matching network is shown as in figure 2. It is worth noting that the quality of the speckle pattern is crucial to the final matching result. In order to obtain reliable matching results, the speckle patterns projected in this paper are designed and evaluated according to our previous work [28].

In figure 2, the whole framework is composed of epipolar rectify, data processing, deep neural network and post-processing. Epipolar rectify is executed first that it helps us effectively simplify a two-dimensional search problem to a one-dimensional matching problem [48]. Then, in data processing, speckle patterns from two perspectives are cropped into a large number of block data with different sizes to input the network during training. Taking the left camera as the main view, for example, for any point (x_{left}, y_{left}) to be matched in the left camera, the data processing module will output a block data (centered on point (x_{left}, y_{left}) with a radius of R). R represents the window radius of block-matching operation. For the 640×480 resolution, $R = 9$ pixels is acceptable and the block size $L \times L$ is 19×19 pixels. For the right camera, the data processing module outputs a block data (centered on point (x_{left}, y_{left}), the left radius is $R + D_{\min}$, the right radius is $R + D_{\max}$ and the vertical radius is R). The meanings of $D_{\min}$ and $D_{\max}$ represent the minimum and maximum disparity values in our stereo vision system and $D(= D_{\max} - D_{\min} + 1)$ in figure 2 is the absolute disparity range, respectively. In other words, by inputting a pair of block data (centered on the point to be matched and its all corresponding candidate points) into the network at the same time, the only thing our network needs to do is to search the correct candidate point within the pre-defined local disparity range. In traditional block-based stereo matching networks, stereo matching is regarded as a binary classification problem [35]. Therefore, for a stereo vision system with a wide epipolar line, there are multiple candidate points for each point to be matched, resulting in a serious imbalance between positive and negative samples in the training dataset (containing one positive sample and multiple negative samples). Different from traditional block-based stereo matching networks, our network treats the whole matching problem of each point as only one sample. Our method is not affected by the imbalance of positive and negative samples so

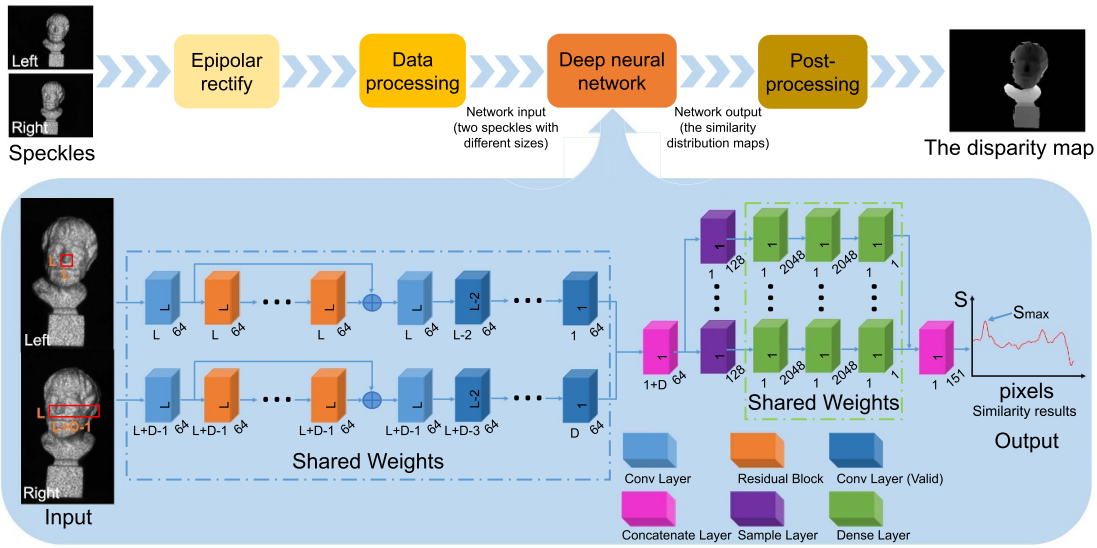

Figure 2. The diagram of the proposed stereo matching method. The whole framework is composed of epipolar rectify, data processing, deep neural network and post-processing. Epipolar rectify is executed first that it helps us effectively simplify a two-dimensional search problem to a one-dimensional matching problem. Then, in data processing, speckle patterns from two perspectives are cropped into a large number of block data with different sizes to input the network during training. Deep neural network, consisting of convolutional layers, residual blocks, concatenate layer, sample layer and dense layer, is used to predict the similarity results from the input data (two speckle block data with different sizes). Finally, these similarity results of all valid points obtained are collected and further processed to yield the disparity map with sub-precision during post-processing.

that our stereo matching networks can more focus on building a matching function with higher accuracy, which makes it possible to obtain a high-precision disparity map.

To realize the high performance of stereo matching above, we adopt the Siamese structure as the matching strategy of the proposed network. The Siamese network will join two shared-weight sub-networks to the front-end section of the network so that input data from different sub-networks are processed uniformly to extract the feature tensor with the same number of dimensions [35, 36, 49]. Specifically speaking, for each shared-weight sub-networks, there are mainly three different layers and blocks including convolutional layers, residual blocks and convolutional layers with valid padding. Convolutional layers can extract different levels of feature information from the input data after training. As the network becomes deeper, it means that the different levels of features extracted are richer, more abstract and more semantic information. However, optimizing such a deeper network is nontrivial due to the degradation problem. Therefore, in our shared-weight sub-networks, we introduce some residual blocks to increase convolutional layers with considerable depth, which can improve the network's feature extraction capability and network performance. Then, some convolutional layers with valid padding are not only exploited to extract and purify features of input tensors to reduce the size of tensors, but also expand the receptive field of each pixel of output tensors to overlap more area of corresponding input block data. The receptive field, which represents the region of the input image that affects any pixel of the output tensor after a series of convolutional layers or pooling layers as shown in figure 3(a), can be described as in theory:

$$RF_k = RF_{k-1} + \left[(f_k - 1) \times \prod_{i=0}^{k-1} s_i \right], \tag{9}$$

where RF_k and RF_{k-1} are the receptive field of each pixel of output tensors at layer k and $k-1$, f_k the kernel size of filters at layer k, s_i the strides of convolutional operation at layer k. In our whole network, $f_k = 3 \times 3$ and $s_i = 1$ are adopted in all convolutional layers and residual blocks. In many computer vision tasks, especially in dense prediction tasks such as semantic image segmentation, stereo vision and optical flow estimation which require pixel-wise prediction for the input image, it is important for each pixel outputted by the network to have a large receptive field so that the network does not ignore any important information during prediction [50].

Specifically, for the input block data with the size of 19×19, it only needs to continuously use nine convolutional layers with valid padding to make the features extracted by the network gather on the central pixel to be matched and the receptive field of which pixel just enough overlaps the entire input data theoretically. However, in fact, the receptive field of CNNs is much smaller than its theoretical value,

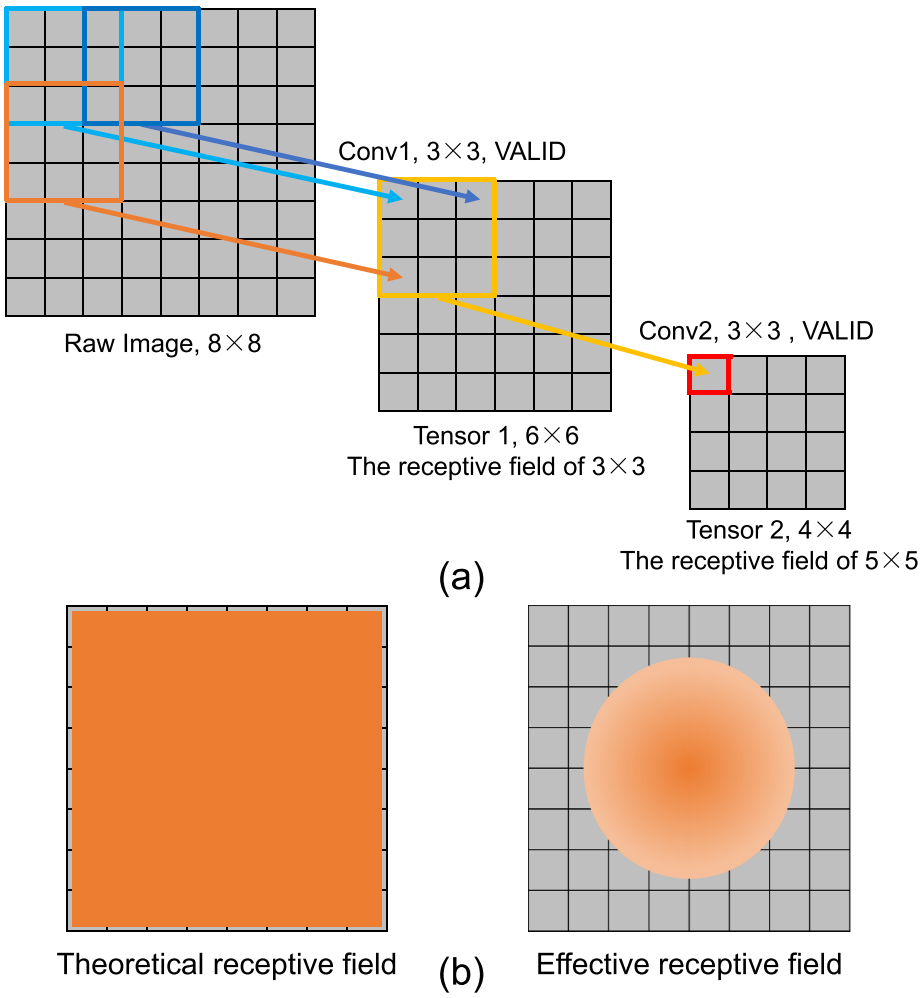

Figure 3. (a) The broaden operation of the receptive field. (b) The diagram of the theoretical receptive field and the effective receptive field.

especially for networks with deeper layers [51, 52]. It is worth noting that the concept of effective receptive fields (ERF) is proposed and can take responsibility for this interesting phenomenon [52]. After analyzed and discussed the receptive field for deep CNNs, it reveals that not all pixels in the receiving field contribute equally to the response of the output unit. In the forward propagation of the networks, the center pixel in the receptive field can propagate information to the output through more different paths, while the pixels in the edge area of the receiving field can propagate their impacts by only a few paths. In backward propagation, the gradient from the output unit will propagate on all paths so that the central pixel has a much larger effect on the output. In addition, it has been proved that the influence of the receptive field is Gaussian distribution in many cases, which means ERF occupies only a small part of the theoretical receptive field and rapidly decays from the center in figure 3(b).

Based on the above analysis, in our sub-network, except for nine convolutional layers with valid padding according to Luo's work [36], some additional but necessary convolutional layers and residual blocks are stacked at the head to further broaden ERF in figure 2. Benefit from the shared-weight mechanism and ERF, block data centered on the points to be matched and their candidate points are treated independently to implement feature extraction and obtain the feature tensors with the 64 dimensions. The idea of this operation is the same as the traditional non-learning matching method, but the difference is that our proposed method has more auto-learning parameters, thereby enable improving network performance and enhancing the generalization ability of the network.

After the feature tensors of the points to be matched and their candidate points are generated in shared-weight sub-networks, as an important operation without learnable weights, the concatenate layer is applied for the feature combination on channel axis. Next, the network needs to estimate the similarity measurement between the point to be matched and each of its candidate points, which can be expressed as the distance between its feature vectors. To address this issue, several similarity measurement methods (e.g. Inner product, Euclidean distance, Hamming distance and Information entropy) have been proposed, which

can provide the similarity results from the input data. In our method, a learnable similarity measurement method is proposed to accurately evaluate the similarity between the feature tensors using dense layers. Inspired by the shared-weight mechanism, dense layers will independently estimate the similarity measurement between the point and each of its candidate points. As the customized layer, sample layers are used to extract two feature tensors after the concatenate layer and merge them on the channel axis as the input of the dense layer. And then, D feature tensors are processed uniformly in these shared-weight dense layers to obtain the final similarity results using the second concatenate layer. During training, we used SGD to minimize the cross-entropy loss with softmax, thereby updating the weights that parameterize the network. Furthermore, for our matching networks, the special ground truth for the points to be matched is a smooth target distribution $P_{gt}(i)$ centered around its correct match point i_{gt} by the following equation:

$$P_{gt}(i) = \begin{cases} 0.5, & if\ i = i_{gt}, \\ 0.2, & if\ |i - i_{gt}| = 1, \\ 0.05, & if\ |i - i_{gt}| = 2, \\ 0, & otherwise. \end{cases} \tag{10}$$

This 3-pixel error metric is smooth to relieve the sample imbalance in the training dataset containing only one positive sample, which helps realize the robust matching with high performance. Here we have done the whole stereo matching for each point $p(x, y)$ and obtain its similarity results $C(p, d)$ (that is the initial matching cost) over all possible candidates.

In post-processing, it is firstly performed based on cost aggregation that the initial matching cost $C(p, d)$ can be further optimized using the SGM method [53]. It is noteworthy that in SGM the initial matching cost predicted by the network should be additionally processed by the negative operation, which means that the correct cost value of the point to be matched should be the minimum value of the matching cost. Finally, the aggregated matching cost $S(p, d)$ will be used for directly obtaining the disparity with the lowest cost by winner-takes-all and enhancing the precision of stereo matching by subpixel disparity refinement, which can achieve more accurate matching results by a five-point quadratic curve fitting model:

$$D_{sub}(p) = d_{int} - \frac{S(p, d_{int} + 1) - S(p, d_{int} - 1)}{2[S(p, d_{int} + 1) + S(p, d_{int} - 1) - 2S(p, d_{int})]}, \tag{11}$$

where $d_{int} = D_{int}(p)$ and $D_{sub}(p)$ represent the correct disparity with pixel precision and subpixel precision. There are followed by some simple post-processing operations are adopted to obtain a dense disparity map. The left-right consistency check is applied for invalidating occlusions and mismatches. The removal operation of noise peaks is performed through a 4-connected segment algorithm [53]. Although our matching network can obtain a relatively accurate disparity map, it still cannot meet the needs of high-precision measurement. Next, the phase information obtained by one high-frequency fringe pattern will be used to refine the disparity results with subpixel precision. Furthermore, the new ground truth for the matching network should be adjusted by combining the five-point quadratic curve fitting and equation (10) in future work.

2.3. Disparity optimization

The 3D geometry retrieved directly from the disparity maps from the speckles has low precision shown in figure 4. To improve the 3D reconstructed precision, high-precision wrapped phases in section 2.1 are exploited to refine the initial disparity map to acquire accuracy disparity results. In disparity optimization, the coarse disparity maps from the speckles provide a rough matching location for each valid point and the only thing we need to do is to search the correct candidate point within the narrow range based on the known high-precision phase prior. In our method, stereo matching based on the phase information is implemented by minimizing the difference between wrapped phases and the integer pixel optimization is realized by the following operation [26, 30]:

$$Diff^{\phi}(i) = \phi_L(x, y) - \phi_R(x + d_{sp} + i, y), \tag{12}$$

$$Diff^{\phi}_{\min}(i) = \min\left[Diff^{\phi}(i), 2\pi - Diff^{\phi}(i)\right], \tag{13}$$

where $i = 0, \pm1, \pm2, \cdots, \pm5$ and $d_{sp} = round(D_{sub}(p))$. Since the wrapped phase can be used instead of the absolute phase, it needs to be considered about the range of the wrapped phase that an exact phase difference

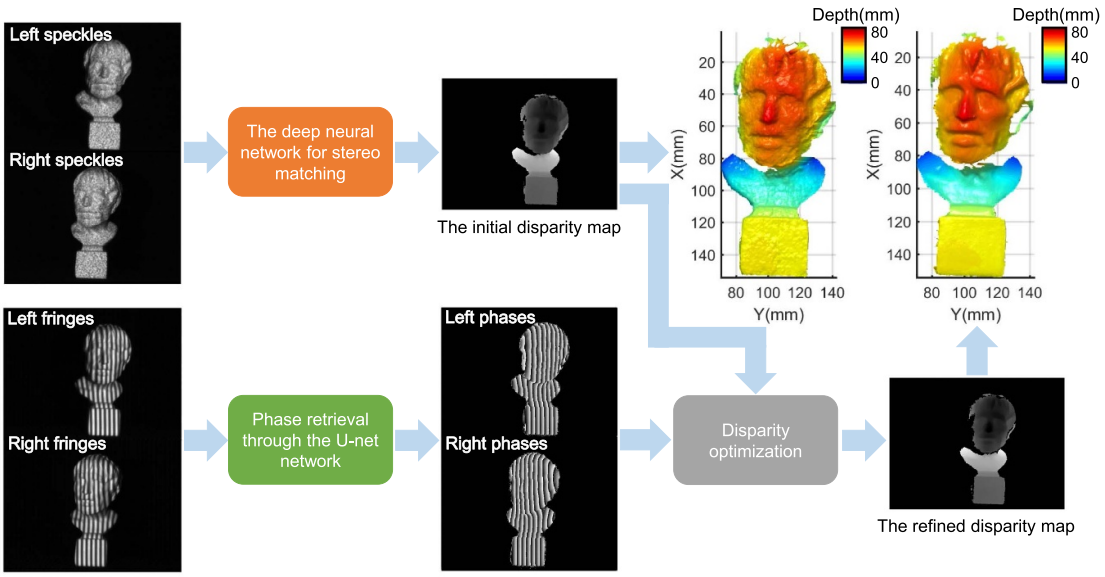

Figure 4. The diagram of the proposed composite deep learning framework for absolute 3D shape measurement based on single fringe phase retrieval and speckle correlation.

is obtained in equation (13). By minimizing the difference between wrapped phases, the best matching point can be found:

$$Diff^{\phi}_{\min}(d_{\min}) = \min_{i} Diff^{\phi}_{\min}(i), \tag{14}$$

$$D^{\phi}_{int} = d_{sp} + d_{\min}, \tag{15}$$

where the disparity D^{ϕ}_{int} is pixel-to-pixel correspondence between two camera views. Then, it must be refined to achieve the subpixel disparity optimization:

$$D^{\phi}_{sub} = D^{\phi}_{int} + \begin{cases} \frac{\phi_L(x,y)-\phi_R(x+D^{\phi}_{int},y)}{\phi_R(x+D^{\phi}_{int}+1,y)-\phi_R(x+D^{\phi}_{int},y)}, & \phi_L(x,y)-\phi_R(x+D^{\phi}_{int},y) > 0, \\ \frac{\phi_L(x,y)-\phi_R(x+D^{\phi}_{int},y)}{\phi_R(x+D^{\phi}_{int},y)-\phi_R(x+D^{\phi}_{int}-1,y)}, & \phi_L(x,y)-\phi_R(x+D^{\phi}_{int},y) < 0. \end{cases} \tag{16}$$

This disparity optimization method can guarantee accuracy stereo matching and obtain dense disparity maps with subpixel precision in figure 4.

3. Experiments

To verify the actual performance of the proposed 3D shape measurement method, a common FPP system is set up including two monochrome camera (Basler acA640-750um with the resolution of 640 × 480) and a DLP projector (LightCrafter 4500Pro with the resolution of 912 × 1140). In our system, the baseline between the stereo cameras is about 270 mm. To measure objects with a depth range of −100 mm to 100 mm, the corresponding disparity constraints are limited to −90 to 60 pixels. For the optimized speckle patterns in this system, the designed parameters are set to include $N_r = 16$ pixels (i.e. the wavelength of fringes is 16 pixels), $M_r = 32$ pixels, $M_s = 1$ pixels and the speckle area of each sub-window is at about 40% [28].

In our experiment, we collected the dataset for training, validation and testing from 1200 different scenes including the random combination of 30 simple and complex objects. The whole dataset has 1200 image pairs, which are divided into 800 image pairs for training, 200 image pairs for validation and 200 image pairs for testing. In supervised learning, the use of high-quality datasets, including input data and ground truth, is very important for learning-based methods. To prepare high-quality ground truth for our network, the 12-step phase-shifting fringe patterns with different frequencies (including 1, 8 and 57) are sequentially projected on the surfaces of multiple samples and synchronously captured by the camera. To monitor during training the accuracy of the neural networks on data that they have never seen before, the scenes in these training, verification and testing datasets are separate from each other. The captured sample data are demonstrated in figure 5. The first to the fourth row shows the captured fringe images with different

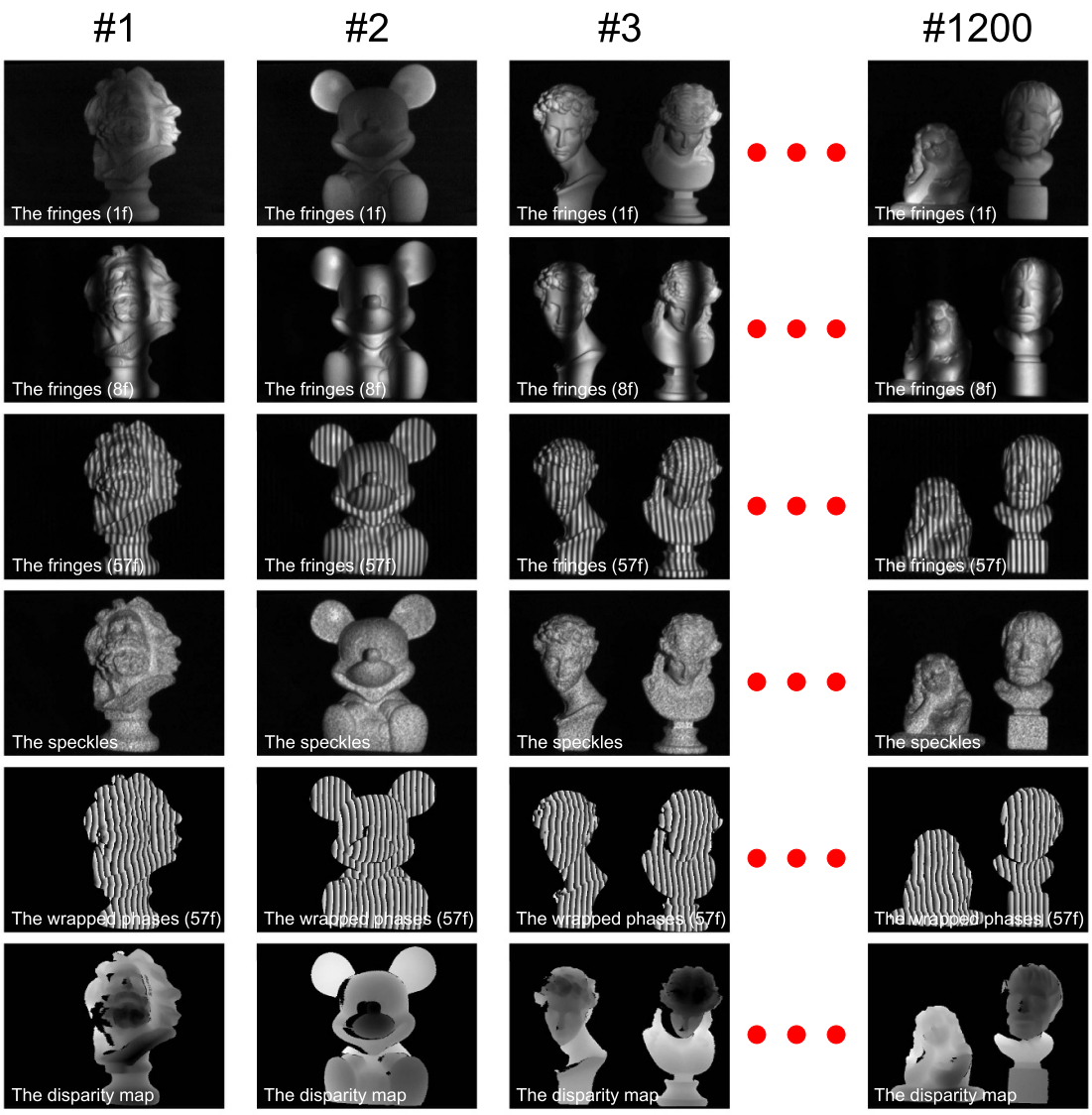

Figure 5. The dataset for training, validation and testing from 1200 different scenes including the random combination of 30 simple and complex objects.

frequencies (including 1, 8, 57) and the captured speckles images, respectively. By combining PSP [12] and multi-frequency temporal phase unwrapping techniques [18], high-precision absolute phase maps with high completeness can be generated and exploited to obtain dense disparity maps with subpixel precision as ground truth of the networks by robust phase matching according to section 2.3. The results are shown in the fifth to the sixth row of figure 5. It is noted that before being fed into the networks, the raw fringe images and speckles images were divided by 255 for normalization, which can make the learning process easier for the network. Moreover, for a preferable selection of training objects, one is suggested choosing objects without very dark or shiny surfaces to ensure captured fringe images and speckles images with enough signal-to-noise ratio or without saturated points.

During the training of the network, for the proposed phase retrieval network, the loss function is set as mean square error (*MSE*), the optimizer is *Adam*, the size of mini-batch is 2 and the training epoch is set as 300. For the proposed stereo matching network, since block matching is a pixel-by-pixel matching method, 18 500 000 valid points are randomly selected from 800 training samples as specific training data and 3 350 000 valid points are randomly selected from 200 validation samples as new validation data. The size of batch is 128, the loss function is set as the cross-entropy loss with softmax and the training iteration steps is set as 40 000. Different from the training procedure, we can reduce the time cost of the proposed stereo matching method during testing by followed Luo's proposal [36]. For feature extraction, our siamese network computes a 64-dimensional feature representation for every pixel. To efficiently obtain the cost volume, we compute the 64-dimensional feature representation only once for every pixel and during computation of the cost volume, we re-use its 64-dimensional feature representation for all disparities that involve this location. In this way, our stereo matching network can obtain a relatively accurate matching result in 0.46 s.

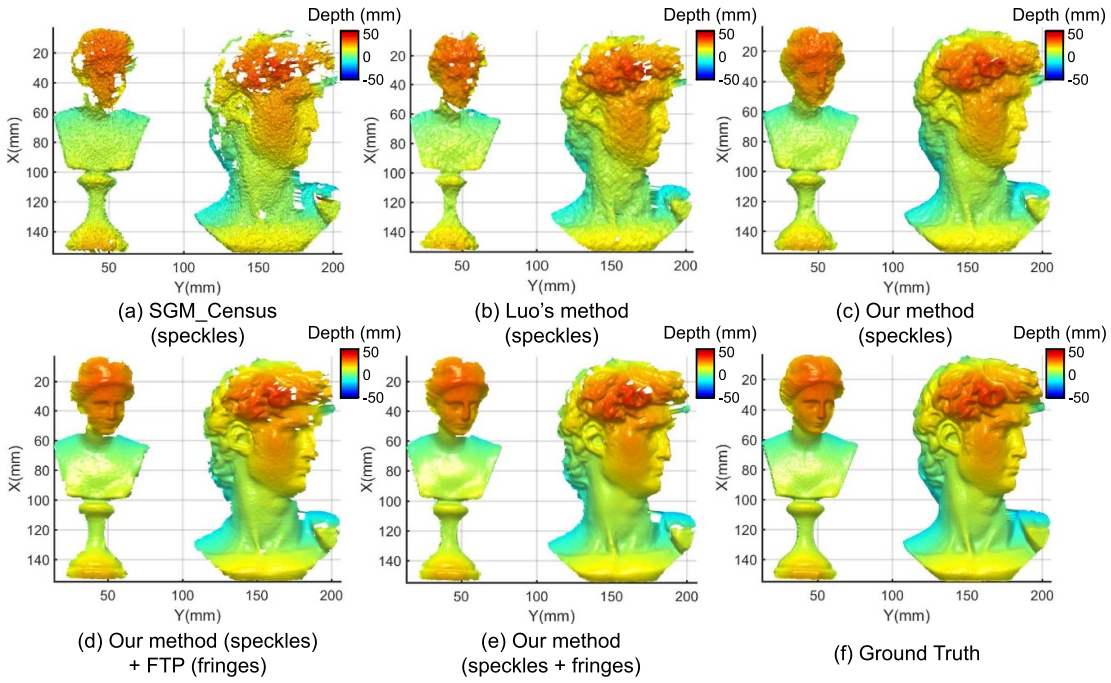

Figure 6. Comparison of the 3D reconstruction results for a representative sample on the testing dataset using different methods. (a) SGM Census (speckles), (b) Luo's method (speckles), (c) our method (speckles), (d) our method (speckles) and FTP (fringes), (e) our method (speckles + fringes) and (f) ground truth.

3.1. Experimental comparison of different methods

In the first experiment, to reveal the high performance of the proposed method, the trained models are utilized to make predictions on the testing dataset, the traditional SGM Census [54], Luo's method [36] and FTP are also implemented for comparison in figure 6. For experimental comparison with stereo matching methods including SGM Census and Luo's method, the census transform and Luo's network with the same block size of 19×19 are used to calculate the initial matching cost, which are both performed by same post-processing in section 2.2. From figure 6(a), there are some mismatch areas and low-precision 3D reconstruction results using SGM Census, which proved that this non-parametric matching method is difficult to provide reliable high-precision matching results for the stereo vision systems with a wide baseline. Different from Luo's method, except for nine convolutional layers with valid padding, some additional but necessary convolutional layers and residual blocks are stacked at the head of the network to further enhance the ability of feature extraction. Besides, the fully connected layers with shared weights are used instead of the original inner product to improve the accuracy of the network's similarity measurement. It is easy to conclude based on these optimization strategies that our matching network can output more accurate dense disparity results shown in figures 6(b)–(c).

For the comparison between FTP and our phase retrieval method, the wrapped phases generated using these two methods both are used to refine the disparity results obtained based on our matching network. It can be found in figure 6(d) that the 3D reconstructed result from FTP has many local details with severe distortion and blurred surfaces due to the inevitable spectral leakage and overlapping in the frequency domain. Compared with FTP, the high-quality 3D reconstruction result in figure 6(e) is yielded by our phase retrieval method through the U-net network, which can automatically utilize both the low-level and high-level feature information of fringes at the same time to optimize its performance of the phase extraction. Among these 3D reconstruction results, the final 3D result produced by our proposed method is almost visually reproduced the ground truth in figure 6(f), so it proves that the proposed 3D shape measurement method can achieve high-efficient, high-precision and robust 3D measurement.

3.2. Quantitative analysis of 3D reconstruction accuracy

Next, the 3D measurement accuracy that our system can achieve will be verified by measuring a standard ceramic plate and a pair of standard ceramic spheres with the diameter of 50.8 mm. Figures 7(a), (c), (e), (g), (i) and (k) displays the related 3D reconstruction results. To obtain the 3D measured errors of the plane, the plane fitting is performed using the 3D reconstruction data to generate the fitted planes as the ground truth. Likewise, for the 3D measurement of the standard ceramic spheres with curve shape, the sphere fitting is used to obtain the correct measurement error. Then, the differences between the measured data and the

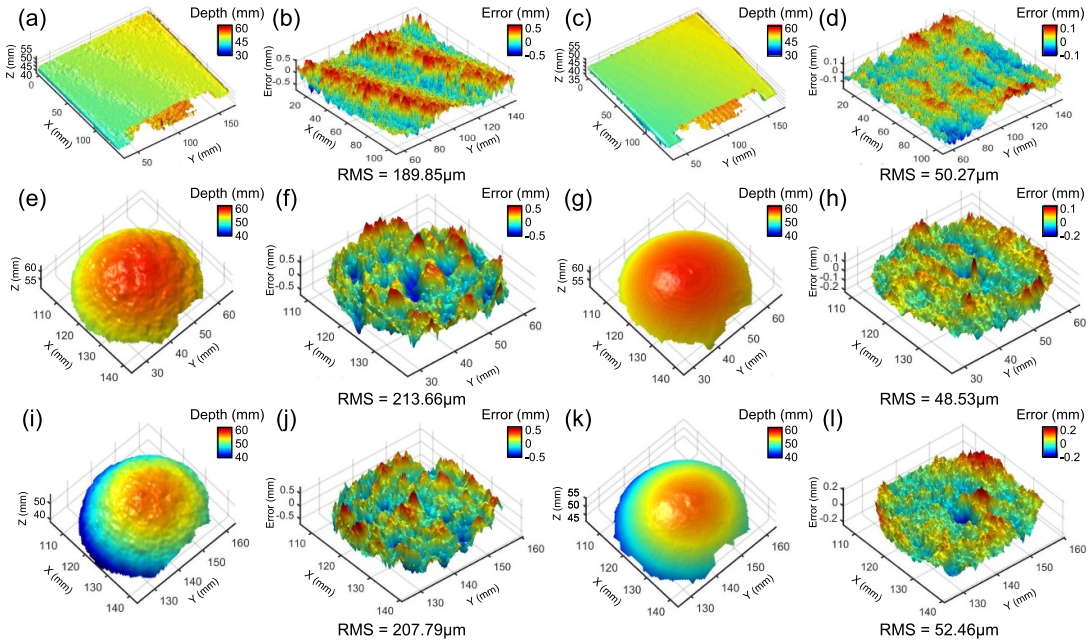

Figure 7. Precision analysis for measuring the plane and a pair of standard ceramic spheres. (a) The 3D reconstruction results of the plate using our method (speckles), (b) the distribution of the errors of (a), (c) the 3D reconstruction results of the plate using our method (speckles + fringes), (d) the distribution of the errors of (c), (e) the 3D reconstruction results of the sphere A using our method (speckles), (f) the distribution of the errors of (e), (g) the 3D reconstruction results of the sphere A using our method (speckles + fringes), (h) the distribution of the errors of (g), (i) the 3D reconstruction results of the sphere B using our method (speckles), (j) the distribution of the errors of (i), (k) the 3D reconstruction results of the sphere B using our method (speckles + fringes), (l) the distribution of the errors of (k).

ground truth are shown in figures 7(b), (d), (f), (h), (j) and (l). It can be found from these distributions of the errors that our method based on the speckles can only obtain these dense but slightly rough 3D measurement results with the precision of about 200 μm. Then, the RMS of the 3D measurement accuracy can be increased to 50 μm by using additional phase information to refine the measurement results. These experiments prove once again that our method can achieve high-precision 3D reconstruction results.

4. Discussions and conclusion

In summary, we have presented a composite deep learning framework for absolute 3D shape measurement based on single fringe phase retrieval and speckle correlation. First, an improved U-net network, which is used for phase retrieval from one 57-period fringe image, is proposed to output the wrapped phases without background information. Compared with FTP, the high-quality wrapped phases are yielded by the U-net network, which can automatically utilize both the low-level and high-level feature information of fringes at the same time to optimize its performance of the phase extraction. Since the trained network is only focused on recovering the phase information in the foreground, it enables our method to generate a more accurate phase. On the other hand, we adopt deep neural networks to beat the stereo matching problem for FPP, which substantially promotes the matching reliability and efficiency compared with the state-of-the-art stereo matching methods. Specifically, to realize the high performance of stereo matching, we adopt the Siamese structure as the matching strategy of the stereo matching network. Different from Luo's method, except for nine convolutional layers with valid padding, some additional but necessary convolutional layers and residual blocks are stacked at the head of the network to further enhance the ability of feature extraction. Besides, the fully connected layers with shared weights are used instead of the original inner product to improve the accuracy of the network's similarity measurement. It can be proved based on these optimization strategies that our matching network can output more accurate and dense 3D measurement results with a precision of about 200 μm. Finally, the wrapped phases predicted by the phase retrieval network as the enhanced phase constraint can be exploited to refine the initial disparity map to dense and accuracy disparity maps for absolute 3D measurement without phase unwrapping. Experimental results have demonstrated the success of our deep-learning-based method using only two projection patterns in its ability to produce dense and accuracy disparity maps to realize high-precision absolute 3D shape measurement with an accuracy of 50 μm for objects with complex surfaces.

In the future, it should be also mentioned that there are several aspects that exist that need to be solved and improved in our deep-learning-based 3D measurement system, which we will leave for future consideration. First, it is worth noting that the computational overhead of our proposed stereo matching network is quite expensive due to block-matching and complex post-processing. Therefore, in order to further improve the efficiency of stereo matching, an end-to-end stereo matching network should be implemented as carefully as possible for FPP. Second, it is not difficult to find that the measurement accuracy achieved by the current stereo matching network cannot meet the requirements of some precision measurement applications without the enhanced phase constraint. How to handle this situation is another interesting direction for further investigation. For example, the new ground truth for the matching network should be adjusted by combining the five-point quadratic curve fitting and the 3-pixel error metric. At last, different from traditional non-learning methods, it needs to be known that the robustness of learning methods for measuring different objects should be paid more attention to enable more reliable 3D shape measurement. So, we will develop alternative techniques to design a more efficient 3D reconstruction method for high-speed and high-precision 3D shape measurement system using deep learning.

Acknowledgment

This work was supported by National Natural Science Foundation of China (61722506, 61705105), National Key R&D Program of China (2017YFF0106403), Leading Technology of Jiangsu Basic Research Plan (BK20192003), Final Assembly '13th Five-Year Plan' Advanced Research Project of China (30102070102), Equipment Advanced Research Fund of China (61404150202), The Key Research and Development Program of Jiangsu Province (BE2017162), Outstanding Youth Foundation of Jiangsu Province (BK20170034), National Defense Science and Technology Foundation of China (0106173), '333 Engineering' Research Project of Jiangsu Province (BRA2016407), Fundamental Research Funds for the Central Universities (30917011204, 30919011222) and Open Research Fund of Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense (3091801410411).

ORCID iDs

Wei Yin https://orcid.org/0000-0002-9148-3401
Jinxin Zhong https://orcid.org/0000-0003-1522-0690
Shijie Feng https://orcid.org/0000-0001-8261-5276
Tianyang Tao https://orcid.org/0000-0003-1122-5785
Jing Han https://orcid.org/0000-0002-1033-566X
Lei Huang https://orcid.org/0000-0002-5645-5173
Qian Chen https://orcid.org/0000-0002-1909-302X
Chao Zuo https://orcid.org/0000-0002-1461-0032

References

[1] Gorthi S S and Rastogi P 2010 Fringe projection techniques: whither we are? *Opt. Laser Eng.* **48** 133–40
[2] Feng S, Zhang L, Zuo C, Tao T, Chen Q and Guohua G 2018 High dynamic range 3d measurements with fringe projection profilometry: a review *Mea. Sci. Technol.* **29** 122001
[3] Zhang Z 2012 Review of single-shot 3d shape measurement by phase calculation-based fringe projection techniques *Opt. Laser Eng.* **50** 1097–106
[4] Zhang S 2018 Absolute phase retrieval methods for digital fringe projection profilometry: A review *Opt. Laser Eng.* **107** 28–37
[5] Liu X, Cai Z, Yin Y, Jiang H, Dong H, Wenqi H, Zhang Z and Peng X 2017 Calibration of fringe projection profilometry using an inaccurate 2d reference target *Opt. Laser Eng.* **89** 131–7
[6] Yin W, Feng S, Tao T, Huang L, Zhang S, Chen Q and Zuo C 2019 Calibration method for panoramic 3d shape measurement with plane mirrors *Opt. Express* **27** 36538–50
[7] Zhang S 2018 High-speed 3d shape measurement with structured light methods: A review *Opt. Laser Eng.* **106** 119–31
[8] Feng S, Zuo C, Tao T, Yan H, Zhang M, Chen Q and Guohua G 2018 Robust dynamic 3-d measurements with motion-compensated phase-shifting profilometry *Opt. Laser Eng.* **103** 127–38
[9] Zhang Z, Towers C E and Towers D P 2006 Time efficient color fringe projection system for 3d shape and color using optimum 3-frequency selection *Opt. Express* **14** 6444–55
[10] Zhang Q and Xianyu S 2005 High-speed optical measurement for the drumhead vibration *Opt. Express* **13** 3110–16
[11] Xianyu S and Chen W 2001 Fourier transform profilometry: a review *Opt. Laser Eng.* **35** 263–84
[12] Zuo C, Feng S, Huang L, Tao T, Yin W and Chen Q 2018 Phase shifting algorithms for fringe projection profilometry: A review *Opt. Laser Eng.* **109** 23–59
[13] Xianyu S and Zhang Q 2010 Dynamic 3-d shape measurement method: a review *Opt. Laser Eng.* **48** 191–204
[14] Feng S, Chen Q, Guohua G, Tao T, Zhang L, Yan H, Yin W and Zuo C 2019 Fringe pattern analysis using deep learning *Adv. Photonics* **1** 025001
[15] Feng S, Zuo C, Yin W, Guohua G and Chen Q 2019 Micro deep learning profilometry for high-speed 3d surface imaging *Opt. Laser Eng.* **121** 416–27

[16] Xianyu S and Chen W 2004 Reliability-guided phase unwrapping algorithm: a review *Opt. Laser Eng.* **42** 245–61
[17] Zhao M, Huang L, Zhang Q, Xianyu S, Asundi A and Kemao Q 2011 Quality-guided phase unwrapping technique: comparison of quality maps and guiding strategies *Appl. Opt.* **50** 6214–24
[18] Zuo C, Huang L, Zhang M, Chen Q and Asundi A 2016 Temporal phase unwrapping algorithms for fringe projection profilometry: A comparative review *Opt. Laser Eng.* **85** 84–103
[19] Wang Y and Zhang S 2012 Novel phase-coding method for absolute phase retrieval *Opt. Lett.* **37** 2067–9
[20] Yin W, Zuo C, Feng S, Tao T, Yan H, Huang L, Jiawei M and Chen Q 2019 High-speed three-dimensional shape measurement using geometry-constraint-based number-theoretical phase unwrapping *Opt. Laser Eng.* **115** 21–31
[21] Liu K, Wang Y, Lau D L, Hao Q and Hassebrook L G 2010 Dual-frequency pattern scheme for high-speed 3-d shape measurement *Opt. Express* **18** 5229–44
[22] Yin W, Chen Q, Feng S, Tao T, Huang L, Trusiak M, Asundi A and Zuo C 2019 Temporal phase unwrapping using deep learning *Sci. Reports* **9** 1–12
[23] Zhong K, Zhongwei Li, Shi Y, Wang C and Lei Y 2013 Fast phase measurement profilometry for arbitrary shape objects without phase unwrapping *Opt. Laser Eng.* **51** 1213–22
[24] Liu X, Yang Y, Tang Q, Cai Z, Peng X, Liu M and Qingquan Li 2018 A method for fast 3d fringe projection measurement without phase unwrapping *Sixth Int. Conf. on Optical and Photonic Engineering (IcOPEN 2018)* vol 10827 Int. Society for Optics and Photonics p 1082713
[25] Cai Z, Liu X, Peng X, Yin Y, Ameng Li, Jiachen W and Gao B Z 2016 Structured light field 3d imaging *Opt. Express* **24** 20324–34
[26] Tao T, Chen Q, Feng S, Yan H, Zhang M and Zuo C 2017 High-precision real-time 3d shape measurement based on a quad-camera system *J. Opt.* **20** 014009
[27] Tao T, Chen Q, Jian D, Feng S, Yan H and Zuo C 2016 Real-time 3-d shape measurement with composite phase-shifting fringes and multi-view system *Opt. Express* **24** 20253–69
[28] Yin W, Feng S, Tao T, Huang L, Trusiak M, Chen Q and Zuo C 2019 High-speed 3d shape measurement using the optimized composite fringe patterns and stereo-assisted structured light system *Opt. Express* **27** 2411–31
[29] Lohry W, Chen V and Zhang S 2014 Absolute three-dimensional shape measurement using coded fringe patterns without phase unwrapping or projector calibration *Opt. Express* **22** 1287–301
[30] Song K, Shaopeng H, Wen X and Yan Y 2016 Fast 3d shape measurement using fourier transform profilometry without phase unwrapping *Opt. Laser Eng.* **84** 74–81
[31] Gai S, Feipeng D and Dai X 2016 Novel 3d measurement system based on speckle and fringe pattern projection *Opt. Express* **24** 17686–97
[32] Jiang C and Zhang S 2017 Absolute phase unwrapping for dual-camera system without embedding statistical features *Opt. Eng.* **56** 094114
[33] Lohry W and Zhang S 2014 High-speed absolute three-dimensional shape measurement using three binary dithered patterns *Opt. Express* **22** 26752–62
[34] LeCun Y, Bengio Y and Hinton G 2015 Deep learning *Nature* **521** 436–44
[35] Zbontar J and Yann L 2015 Computing the stereo matching cost with a convolutional neural network *Proc. of the Conf. on Computer Vision and Pattern Recognition* pp 1592–9
[36] Luo W, Schwing A G and Urtasun R 2016 Efficient deep learning for stereo matching *Proc. of the Conf. on Computer Vision and Pattern Recognition* pp 5695–703
[37] Seki A and Pollefeys M 2017 Sgm-nets: Semi-global matching with neural networks *Proc. of the Conf. on Computer Vision and Pattern Recognition* pp 231–40
[38] Chang J-R and Chen Y-S 2018 Pyramid stereo matching network *Proc. of the Conf. on Computer Vision and Pattern Recognition* pp 5410–18
[39] Liang Z, Feng Y, Guo Y, Liu H, Chen W, Qiao L, Zhou Li and Zhang J 2018 Learning for disparity estimation through feature constancy *Proc. of the Conf. on Computer Vision and Pattern Recognition* pp 2811–20
[40] Liang Z, Guo Y, Feng Y, Chen W, Qiao L, Zhou Li, Zhang J and Liu H 2019 Stereo matching using multi-level cost volume and multi-scale feature constancy *IEEE Transactions on Pattern Analysis and Machine Intelligence*
[41] Vogel C, Schindler K and Roth S 2015 3d scene flow estimation with a piecewise rigid scene model *Int. J. Comput. Vis.* **115** 1–28
[42] Jia X, Ranftl Re and Koltun V 2017 Accurate optical flow via direct cost volume processing *Proc. of the Conf. on Computer Vision and Pattern Recognition* pp 1289–97
[43] Yang Z, Sun Y, Liu S, Shen X and Jia J 2019 Std: Sparse-to-dense 3d object detector for point cloud *Proc. of the Conf. on Computer Vision and Pattern Recognition* pp 1951–60
[44] Chen X, Huimin M, Wan J, Li B and Xia T 2017 Multi-view 3d object detection network for autonomous driving *Proc. of the Conf. on Computer Vision and Pattern Recognition* pp 1907–15
[45] Choi W 2015 Near-online multi-target tracking with aggregated local flow descriptor *Proc. of the Conf. on Computer Vision and Pattern Recognition* pp 3029–37
[46] Ronneberger O, Fischer P and Brox T 2015 U-net: Convolutional networks for biomedical image segmentation *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention* Springer p pages 234–241
[47] Scharstein D and Szeliski R 2002 A taxonomy and evaluation of dense two-frame stereo correspondence algorithms *Int. J. Comput. Vis.* **47** 7–42
[48] Hartley R and Zisserman A 2003 *Multiple View Geometry in Computer Vision* (Cambridge: Cambridge university press)
[49] Bertinetto L, Valmadre J, Henriques J F, Vedaldi A and Torr P H S 2016 Fully-convolutional siamese networks for object tracking *European Conference on Computer Vision* Springer pp 850–65
[50] Zhao H, Shi J, Xiaojuan Q, Wang X and Jia J 2017 Pyramid scene parsing network *Proc. of the Conf. on Computer Vision and Pattern Recognition* pp 2881–90
[51] Zhou B, Khosla A, Lapedriza A, Oliva A and Torralba A 2014 Object detectors emerge in deep scene cnns arXiv:1412.6856
[52] Luo W, Yujia Li, Urtasun R and Zemel R 2016 Understanding the effective receptive field in deep convolutional neural networks *Advances in Neural Information Processing Systems* pp 4898–906
[53] Hirschmuller H 2007 Stereo processing by semiglobal matching and mutual information *IEEE Trans. Pattern Anal. Mach. Intell.* **30** 328–41
[54] Hirschmuller H and Scharstein D 2008 Evaluation of stereo matching costs on images with radiometric differences *IEEE Trans. Pattern Anal. Mach. Intell.* **31** 1582–99