Research Article Vol. 29, No. 9 / 26 April 2021 / Optics Express 13388
Optics EXPRESS

Single-shot 3D shape measurement using an end-to-end stereo matching network for speckle projection profilometry

WEI YIN,[1,2,3,4] YAN HU,[1,2,3,4] SHIJIE FENG,[1,2,3,4,7] LEI HUANG,[5] QIAN KEMAO,[6] QIAN CHEN,[1,2,8] AND CHAO ZUO[1,2,3,4,9,10]

[1] *School of Electronic and Optical Engineering, Nanjing University of Science and Technology, No. 200 Xiaolingwei Street, Nanjing, Jiangsu Province 210094, China*
[2] *Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense, Nanjing, Jiangsu Province 210094, China*
[3] *Smart Computational Imaging (SCI) Laboratory, Nanjing University of Science and Technology, Nanjing, Jiangsu Province 210094, China*
[4] *Smart Computational Imaging Research Institute (SCRI) of Nanjing University of Science and Technology, Nanjing, Jiangsu Province 210019, China*
[5] *Brookhaven National Laboratory, NSLS II 50 Rutherford Drive, Upton, New York 11973-5000, USA*
[6] *School of Computer Science and Engineering, Nanyang Technological University, 639798, Singapore*
[7] *geniusshijie@163.com*
[8] *chenqian@njust.edu.cn*
[9] *zuochao@njust.edu.cn*
[10] *surpasszuo@163.com*

Abstract: Speckle projection profilometry (SPP), which establishes the global correspondences between stereo images by projecting only a single speckle pattern, has the advantage of single-shot 3D reconstruction. Nevertheless, SPP suffers from the low matching accuracy of traditional stereo matching algorithms, which fundamentally limits its 3D measurement accuracy. In this work, we propose a single-shot 3D shape measurement method using an end-to-end stereo matching network for SPP. To build a high-quality SPP dataset for training the network, by combining phase-shifting profilometry (PSP) and temporal phase unwrapping techniques, high-precision absolute phase maps can be obtained to generate accurate and dense disparity maps with high completeness as the ground truth by phase matching. For the architecture of the network, a multi-scale residual subnetwork is first leveraged to synchronously extract compact feature tensors with 1/4 resolution from speckle images for constructing the 4D cost volume. Considering that the cost filtering based on 3D convolution is computationally costly, a lightweight 3D U-net network is proposed to implement efficient 4D cost aggregation. In addition, because the disparity maps in the SPP dataset should have valid values only in the foreground, a simple and fast saliency detection network is integrated to avoid predicting the invalid pixels in the occlusions and background regions, thereby implicitly enhancing the matching accuracy for valid pixels. Experiment results demonstrated that the proposed method improves the matching accuracy by about 50% significantly compared with traditional stereo matching methods. Consequently, our method achieves fast and absolute 3D shape measurement with an accuracy of about 100μm through a single speckle pattern.

1. Introduction

Optical 3D measurements based on structured light projection have become a prevalent non-contact 3D shape measurement technique [1]. With the advantages of simple hardware configuration, high measurement accuracy, and high 3D point cloud density, it has been proven one of the most

#418881 https://doi.org/10.1364/OE.418881
Journal © 2021 Received 30 Dec 2020; revised 24 Feb 2021; accepted 30 Mar 2021; published 16 Apr 2021

promising techniques and is extensively applied in industry inspection and scientific research [2–5]. Essentially, the structured light-based 3D measurement methods can be regarded as an improved form of stereo vision, which is achieved by introducing an additional light source generator (such as a projector) in the system configuration [6]. The light source generator projects a series of specifically coded patterns onto the measured scenes [7]. Compared with stereo vision-based methods, the structured light-based 3D measurement methods can easily overcome the problem of low matching accuracy caused by weak texture regions.

Among the 3D shape measurement methods based on structured light projection, two commonly used structured light patterns are fringe patterns and speckle patterns. Correspondingly, there are two mainstream methods: fringe projection profilometry (FPP) [8–11] and speckle projection profilometry (SPP) [12–14]. In FPP, the projector projects a series of fringe patterns onto the measured scenes. The fringe images modulated by the measured objects are captured synchronously by the camera and then processed to obtain the phase information by using various phase retrieval techniques, such as Fourier transform profilometry (FTP) [15–17] and phase-shifting profilometry (PSP) [18]. However, these methods both adopt the arctangent function which can only provide a wrapped phase with 2π phase jumps. Therefore, it is necessary to perform phase unwrapping to eliminate the phase ambiguity and convert the wrapped phase to the absolute phase [19–25]. To address this issue, several composite phase-shifting schemes (e.g, dual-frequency PSP [26], bi-frequency PSP [27], and 2+2 PSP [28]) have been proposed, which can solve the phase ambiguity problem without significantly increasing the number of projected patterns. However, these methods still require a certain number of projection patterns. As a result, it is difficult to obtain high-precision and absolute phase information from a single fringe image in FPP, which limits its applications in dynamic 3D measurement [29,30].

Different from FPP, the projector in SPP projects a speckle pattern onto the measured scenes. The speckle images modulated by the measured objects are captured synchronously by the stereo camera and then processed to obtain the disparity map by using various stereo matching techniques. The projected speckle pattern designed using a spatial encoding strategy has inherently global uniqueness, which makes the SPP-based 3D measurement methods have the advantage of single-shot 3D reconstruction. Therefore, the key idea of the design method for the speckle pattern is how to ensure that the local speckles are globally unique with respect to the whole projection pattern [31]. These design methods for projection patterns can be grouped into three main classes based on various spatial encoding strategies [7,32,33]: strategies based on non-formal codification [34,35], strategies based on De Bruijn sequences [36–38], and strategies based on M-arrays [39]. In the last few decades, researchers have proposed numerous design methods for the speckles. However, due to the measured objects with complex reflection characteristics and the perspective differences between the stereo camera, it is still difficult to ensure the global uniqueness of each pixel in the whole measurement space by only projecting one speckle pattern [12,14,40], which leads to the common mismatching in actual measurements. In order to solve this problem in SPP, some robust stereo matching algorithms such as SGM [41–43] and ELAS [44] are proposed to acquire dense disparity maps, thus enabling robust absolute 3D measurement. However, these methods achieve reliable stereo matching by smoothing the disparity map, at the cost of matching accuracy. It is easy to understand that projecting multiple speckle images will improve the accuracy of 3D measurement, because more constraints can be exploited to completely guarantee the global uniqueness of the measured scenes. Following this idea, Zhou *etal*. [14] proposed a high-precision 3D surface profile measurement scheme by only projecting a single-shot color binary speckle pattern (CBSP) and a temporal-spatial correlation matching algorithm, which can be applied to measurements of dynamic and static objects. In order to improve the 3D measurement speed, Schaffer *etal*. [12,13] used laser speckles as projected patterns which are switched using an acousto-optical deflector. Its projection rate is more than 10 times higher than the common projection systems. Capturing images of encoded

Optics EXPRESS

objects through two synchronized high-speed cameras, this proposed system achieves high-speed, dense, and accurate 3D measurements of spatially separated objects at 350 frames per second. These proposed SPP methods can achieve high-performance 3D measurement based on speckle projection, but it is impossible to obtain accurate 3D data from a single speckle image. For SPP, it still lacks a stereo matching algorithm using a single speckle pattern that can achieve high-robustness and high-accuracy 3D measurement for the recovery of the fine details of complex surfaces.

Compared with traditional stereo matching methods, recently, many deep learning methods for stereo vision are proposed and have achieved excellent performance of stereo matching [45–52]. There is generally a four-step pipeline for stereo matching, including matching cost calculation, cost aggregation, disparity computation, and disparity refinement, while traditional stereo matching methods perform all four steps using non-learning techniques. Existing learning-based stereo matching methods attempt to exploit deep learning to implement one or multiple of the four steps to obtain better matching results. LeCun *etal*. [45] first adopted the Siamese network to perform block matching for obtaining the initial matching cost and then exploited typical stereo matching procedures, including SGM-based cost aggregation, disparity computation, and disparity refinement to further improve matching results. Luo *etal*. [46] inputted left and right image patches with different sizes into the CNNs for computing the initial matching cost, which will convert the binary classification problem into a multi-classification task, enabling high-efficiency stereo matching. Currently, some end-to-end stereo matching networks have been developed to predict whole disparity maps without post-processing. Kendall *etal*. [49] proposed to generate a 4D cost volume of size $C \times D \times H \times W$ (*i.e.*, $Features \times Disparity \times Height \times Width$) by combining the features of all pixels from the reference image and all candidates among disparity ranges along the epipolar line of the target image. The 4D cost volume is filtered through a series of 3D convolutional layers. The final disparity maps are regressed from the filtered cost volume using a differentiable soft argmin operation, which allows it to achieve matching results with sub-pixel accuracy without any additional post-processing or regularization. Later, Chang *etal*. [51] proposed a pyramid stereo matching network (PSMNet) to further improve the matching accuracy by using the spatial pyramid pooling and multiple hourglass networks based on the 3D CNN. Zhang *etal*. [52] introduced SGM-based cost aggregation and local guided filter into the existing cost aggregation subnetwork to obtain better matching accuracy and the generalization ability of the network.

In this work, we propose a single-shot 3D shape measurement method using an end-to-end stereo matching network for speckle projection profilometry. In supervised learning, the use of high-quality datasets, including input data and ground truth, is very important for learning-based methods. KITTI is a prominent stereo dataset, which promoted the development of deep learning in stereo vision [53]. It is worth noting that KITTI is very challenging because its labels obtained by 3D Lidar are extremely sparse and low-precision. In our method, different from KITTI, by combining 12-step PSP [18] and multi-frequency temporal phase unwrapping techniques [22], high-precision absolute phase maps with high completeness can be obtained to generate dense disparity maps with subpixel precision by phase matching, which will be as the high-quality ground truth for our stereo matching network. For the architecture of our proposed network, a multi-scale residual subnetwork is first leveraged to synchronously extract compact feature tensors with 1/4 resolution from speckle images for constructing the 4D cost volume. Considering that the cost filtering operation using 3D convolutional layers is computationally expensive, a lightweight 3D U-net network is proposed to implemented efficient 4D cost aggregation for achieving higher matching performance. In addition, because the disparity maps (as the ground truth) in the SPP dataset has valid values only in the foreground, a simple and fast saliency detection network is integrated into our end-to-end network to avoid predicting the invalid pixels in the disparity maps including occlusions and backgrounds, thereby implicitly enhancing the matching accuracy for valid pixels. Based on the proposed method, the matching accuracy is improved by about 50% significantly compared with traditional stereo matching methods. The experiment results demonstrated that the proposed method can achieve fast and absolute 3D shape measurement with an accuracy of about 100μm through a single speckle pattern.

2. Principle

In this section, a single-shot 3D shape measurement method using an end-to-end stereo matching network for speckle projection profilometry will be presented. In our method, a speckle pattern and a series of fringe patterns need to be projected by the projector onto the measured scenes and captured synchronously by the stereo camera. The acquired speckle image pair is first processed by epipolar rectification, and then fed directly into the proposed end-to-end stereo matching network to obtain the corresponding disparity map without the background. The disparity map is converted into the final 3D results after disparity-to-height mapping as shown in Fig. 1. It is clear that the projected speckle pattern and the end-to-end stereo matching network together determine the actual 3D measurement performance of the proposed method.

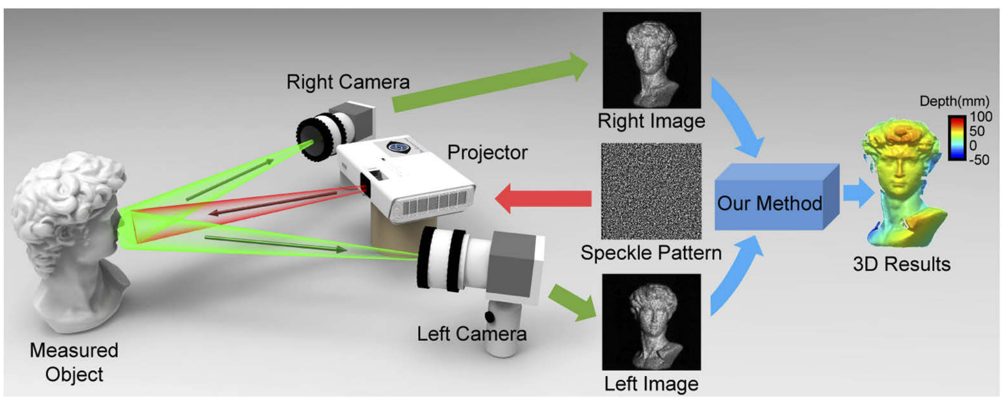

Fig. 1. The diagram of the proposed single-shot 3D shape measurement method using an end-to-end stereo matching network for speckle projection profilometry.

For the speckle pattern, we follow a simple and effective design and evaluation method proposed in our previous work [31]. By introducing epipolar rectification and depth constraint, the only thing the stereo matching algorithms need to do is to search the corresponding pixel within the pre-defined local 1D range rather than the traditional global 2D range, which means that our optimized design method of the speckle pattern just requires the local speckles in the speckle patterns are unique with respect to the local 1D projection space. Based on this idea, the projected speckle pattern is designed and evaluated to assist in improving the 3D measurement performance.

For the proposed end-to-end stereo matching network, there are two aspects that affect its final stereo matching performance. First, for the deep learning-based network approach, the datasets, including input data and ground truth, are very important to efficiently train the stereo matching network. In our method, a series of acquired fringe images are used to generated dense disparity maps with subpixel precision as the high-quality ground truth for our SPP datasets, which potentially determines the trained network's highest matching accuracy and robustness when measuring objects with complex surfaces. In the next subsection, we will discuss in detail how to construct a high-quality SPP dataset using phase-shifting methods and multi-frequency temporal phase unwrapping techniques in FPP. Secondly, for the architecture of our proposed network, although a large number of high-performance learning-based stereo matching networks exist, these networks are generally trained and validated on the KITTI stereo dataset and cannot be directly applied to SPP. KITTI is a prominent stereo dataset, which promoted the development of deep learning in stereo vision [53]. It is worth noting that KITTI is very challenging because its labels obtained by 3D Lidar are extremely sparse and low-precision. Specifically, KITTI is a dataset in the field of autonomous driving, in which the data has the properties of large scale and sparse texture, and its 3D reconstruction accuracy is millimeter precision. In contrast, our stereo matching network aims to achieve high-precision and robust 3D measurements with micron-level accuracy by matching the objects with strong speckle texture information. The specific structure of the proposed network will be presented in detail according to Section 2.2.

2.1. High-quality SPP dataset constructed by using FPP

To build a high-quality SPP dataset, fringe projection profilometry (FPP) is used to obtain high-precision and dense disparity maps as the ground truth. In a common FPP system, there are three main processing steps in FPP: phase extraction, phase unwrapping, and phase-to-height mapping. During phase recovery, sinusoidal fringe-based FPP methods are more prevalent to retrieval the wrapped phase using Fourier transform methods in frequency domain [15] or phase-shifting methods in time domain [18]. Fourier transform profilometry (FTP) has the advantage of single-shot phase extraction but suffers from the spectrum overlapping problem. These methods generally produce coarse wrapped phases with low quality, making it difficult to achieve high-precision 3D acquisition. Different from FTP, phase-shifting profilometry (PSP) can realize pixel-wise phase measurements with higher accuracy unaffected by ambient light, but it needs to project at least three fringe patterns to obtain a phase map theoretically.

In this work, the standard 12-step phase-shifting fringe patterns with shift offset of $2\pi/12$ are adopted because it is quite robust to ambient illumination and varying surface properties:

$$I_n^p(x,y) = 0.5 + 0.5\cos(2\pi f x - 2\pi n/12), \tag{1}$$

where $I_n^p(x,y)(n = 0,1,2,\ldots,11)$ represent fringe patterns to be projected, f is the frequency of fringe patterns. Then the fringe images captured by the camera can be described as

$$I_n^c(x,y) = A^c(x,y) + B^c(x,y)\cos(\phi^c(x,y) - 2\pi n/12), \tag{2}$$

where $I_n^c(x,y)$ represent the intensity of captured fringe images, $A^c(x,y)$, $B^c(x,y)$, and $\phi^c(x,y)$ are the average intensity, the intensity modulation, and the phase distribution of the measured object. 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:

$$\phi^c(x,y) = \tan^{-1}\frac{\sum_{n=0}^{11} I_n^c(x,y)\sin(2\pi n/12)}{\sum_{n=0}^{11} I_n^c(x,y)\cos(2\pi n/12)}, \tag{3}$$

$$B^c(x,y) = \frac{2}{12}\sqrt{\left[\sum_{n=0}^{11} I_n^c(x,y)\sin(2\pi n/12)\right]^2 + \left[\sum_{n=0}^{11} I_n^c(x,y)\cos(2\pi n/12)\right]^2}, \tag{4}$$

$$Mask_v^c(x,y) = B^c(x,y)/255 > Thr1, \tag{5}$$

where $Thr1$ is the preset threshold for the tested object, $Mask_v^c(x,y)$ can be used to identify the valid points in the whole image. 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.

Optics EXPRESS

Due to the truncation effect of the arctangent function in Eq. (3), the obtained phase $\phi^c(x, y)$ is wrapped within the range of $(-\pi, \pi]$, and its relationship with $\Phi^c(x, y)$ is:

$$\Phi^c(x, y) = \phi^c(x, y) + 2\pi k^c(x, y), \tag{6}$$

where $k^c(x, y)$ represents the fringe order of $\Phi^c(x, y)$, and its value range is from 0 to $f - 1$.

In our method, multi-frequency temporal phase unwrapping method (MF-TPU) is exploited to obtain $k^c(x, y)$ for each pixel in the phase map accurately. In MF-TPU, the wrapped phase $\phi^c(x, y)$ is unwrapped with the aid of one (or more) additional wrapped phase map with different frequency. For instance, two wrapped phases $\phi_h^c(x, y)$ and $\phi_l^c(x, y)$ are both retrieved from phase-shifting algorithms by using Eq. (3), ranging from $-\pi$ to π. It is easy to find that the two absolute phases $\Phi_h^c(x, y)$ and $\Phi_l^c(x, y)$ corresponding to $\phi_h^c(x, y)$ and $\phi_l^c(x, y)$ have the following relationship:

$$\begin{cases} \Phi_h^c(x, y) = \phi_h^c(x, y) + 2\pi k_h^c(x, y), \\ \Phi_l^c(x, y) = \phi_l^c(x, y) + 2\pi k_l^c(x, y), \\ \Phi_h^c(x, y) = (f_h/f_l)\Phi_l^c(x, y), \end{cases} \tag{7}$$

where f_h and f_l are the frequency of high-frequency fringes and low-frequency fringes. Based on Eq. (7), $k_h^c(x, y)$ can be calculated by the following formula:

$$k_h^c(x, y) = \frac{(f_h/f_l)\Phi_l(x, y) - \phi_h(x, y)}{2\pi}. \tag{8}$$

Since the fringe order $k_h^c(x, y)$ is integer, ranging from 0 to f_h - 1, Eq. (8) can be adapted as

$$k_h^c(x, y) = Round\left[\frac{(f_h/f_l)\Phi_l^c(x, y) - \phi_h^c(x, y)}{2\pi}\right], \tag{9}$$

where $Round()$ is the rounding operation. When f_l is 1, there will be no phase ambiguity so that $\phi_l^c(x, y)$ is inherently an unwrapped phase. Theoretically, for MF-TPU, this single-period phase can be used to directly assist phase unwrapping of $\phi_h^c(x, y)$ with relatively higher frequency. However, the phase unwrapping capability of MF-TPU is greatly constrained due to the influence of noise in practice. For a normal FPP system, MF-TPU can only reliably unwrap the phase with about 16 periods due to the non-negligible noises and other error sources in actual measurement. Thus, it generally exploits multiple (>2) sets of phases with different frequencies to hierarchically unwrap the wrapped phase step by step, and finally arrives at the absolute phase with high frequency instead of only using the phase with a single period. In our method, three wrapped phases with different frequencies (including 1, 8 and 57) are used to obtain high-precision and dense (57-period) absolute phase.

Finally, phase matching based on the phase information is implemented to obtain the disparity map with integer-pixel precision by minimizing the difference between absolute phases from two perspectives:

$$\Delta\Phi(i) = abs(\Phi_L(x, y) - \Phi_R(x + i, y)), \tag{10}$$

$$\Delta\Phi_{min}(D_{int}) = \min_i \Delta\Phi(i), \tag{11}$$

where i is the candidate disparity value locally in our SPP system based on epipolar rectification and depth constraint, the disparity D_{int} represents the pixel-to-pixel correspondence between two camera views. Then, the disparity refinement is realized to obtain the disparity map with subpixel precision by a simple linear interpolation:

$$D_{sub} = D_{int} + \begin{cases} \frac{\Phi_L(x,y)-\Phi_R(x+D_{int},y)}{\Phi_R(x+D_{int}+1,y)-\Phi_R(x+D_{int},y)}, & \Phi_L(x, y) - \Phi_R(x + D_{int}, y)>0, \\ \frac{\Phi_L(x,y)-\Phi_R(x+D_{int},y)}{\Phi_R(x+D_{int},y)-\Phi_R(x+D_{int}-1,y)}, & \Phi_L(x, y) - \Phi_R(x + D_{int}, y)<0. \end{cases} \tag{12}$$

By phase matching, the high-precision and dense disparity map D_{sub} can be obtained as the ground truth of our high-quality SPP dataset in Fig. 2.

Optics EXPRESS

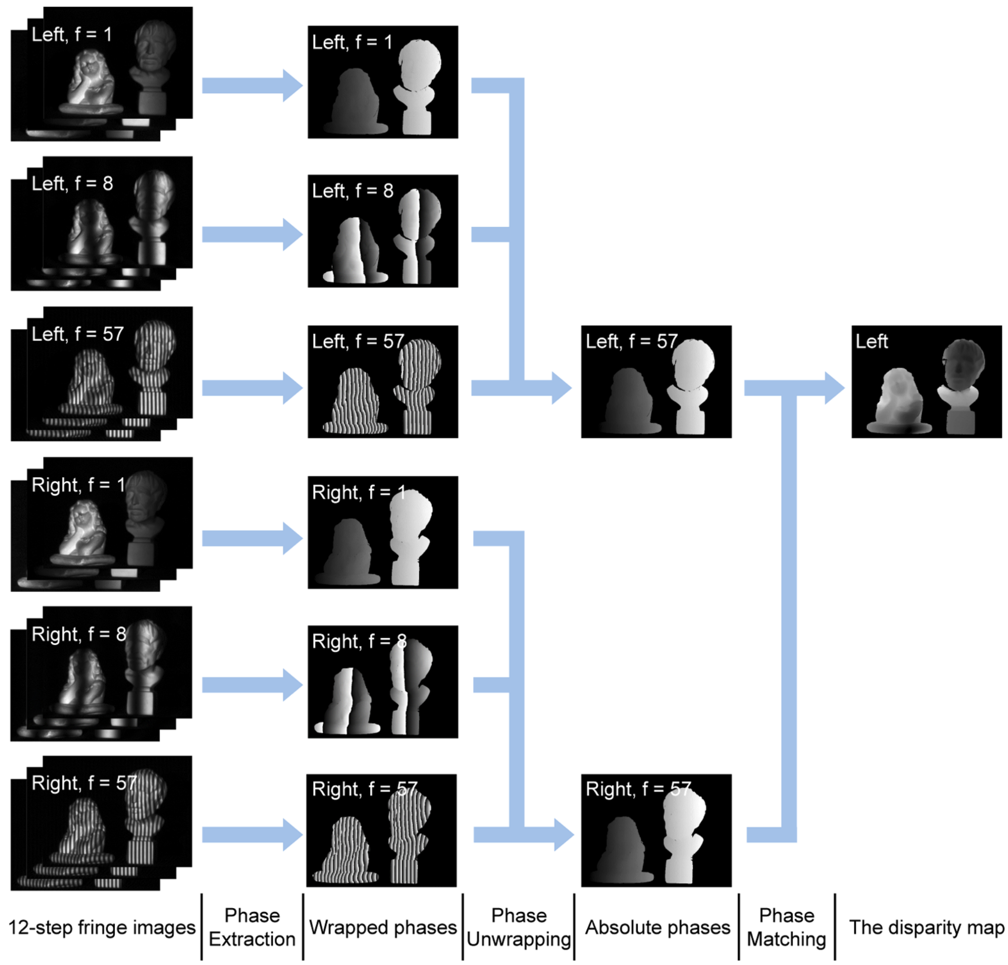

Fig. 2. The diagram of constructing high-quality SPP dataset by using FPP.

2.2. End-to-end stereo matching network

In this subsection, an end-to-end stereo matching network, which is used to solve the stereo matching problem in SPP, is proposed to substantially promote the matching accuracy compared with the state-of-the-art stereo matching methods. Existing high-performance learning-based stereo matching networks are generally trained and validated on the KITTI stereo dataset. In the KITTI stereo dataset, the data has the properties of large scale and sparse texture, and the corresponding 3D reconstruction results have only millimeter precision. In contrast, based on our high-quality SPP dataset, our stereo matching network aims to achieve robust 3D measurements with micron-level accuracy using a speckle image pair. In addition, for the ground truth of our SPP dataset, the disparity map of the sample data has valid values only in the foreground as shown in Fig. 2. Thus, it is difficult to naively exploit these existing end-to-end networks [50–52] to directly obtain the final disparity map, but a simple and fast saliency detection network is integrated into our network to avoid predicting the invalid pixels in the disparity maps including occlusions and backgrounds. Specifically, the schematic diagram of the proposed stereo matching network is shown in Fig. 3.

Optics EXPRESS

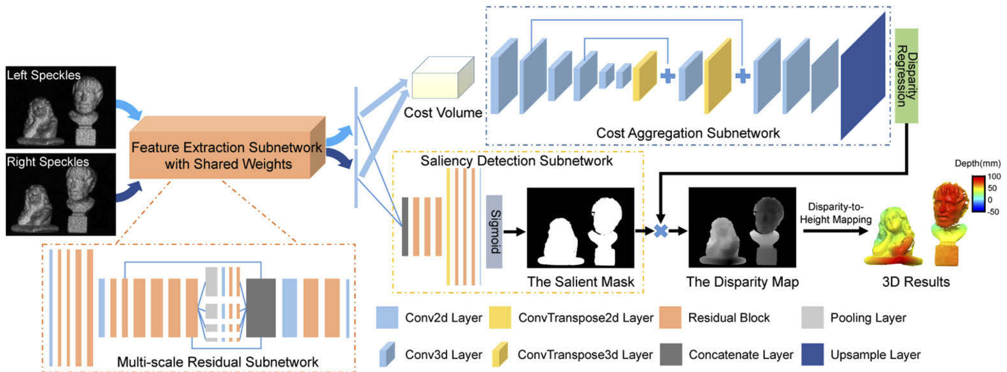

Fig. 3. The schematic diagram of the proposed end-to-end stereo matching network. The whole stereo matching network is composed of a multi-scale residual subnetwork (as the shared feature extraction subnetwork), construction of the 4D cost volume, cost aggregation using 3D convolutional layers, disparity regression, and a saliency detection subnetwork.

In Fig. 3, the whole stereo matching network is composed of a multi-scale residual subnetwork (as the shared feature extraction subnetwork), construction of the 4D cost volume, cost aggregation using 3D convolutional layers, disparity regression, and a saliency detection subnetwork. It is worth noting that before stereo matching epipolar rectification is first executed to simplify the two-dimensional search problem to a one-dimensional matching problem [54]. Then, in feature extraction for matching cost calculation, different from the traditional methods that directly exploit the gray information or color value of the pixel for correspondence matching, our purpose is to calculate the feature representation of each pixel to be matched for the subsequent matching process. Specifically, learning-based methods usually implement feature extraction on the input stereo images simultaneously to obtain rich feature information, which is used to construct a 4D cost volume as the initial matching cost. Therefore, the initial matching accuracy corresponding to the initial matching cost strongly depends on the quality of the extracted feature information.

For the feature extraction subnetwork in our work, a multi-scale residual network is proposed to process the input stereo image pair to obtain rich multi-scale feature information. In this subnetwork, speckle images are first processed by a 2D convolution layer and four residual blocks to obtain 64-channel feature tensors. Considering that the high-resolution matching costs in the subsequent cost aggregation will consume a lot of computational overhead and take up expensive GPU memories, it is necessary to perform a 1/4 downsample operation on the feature tensors. It is worth noting that the extraction of low-resolution feature tensors is not so much a compromise to the expensive computational cost but to keep the feature tensors more compact and achieve high-efficiency feature extraction. Then, the low-resolution feature tensors successively go through six residual blocks for further expanding the receptive field of each pixel of output tensors. It is crucially important that each pixel of feature tensors yielded by the network must have a larger receptive field so that the network will not ignore any important feature information during the prediction period [55]. And then, the multi-scale pooling layers are introduced to downsample the input tensors by 1/4, 1/16, 1/64, and 1/256, which can further compress and extract the main features of the tensors to reduce computation complexity and prevent over-fitting. For these four downsample paths, the feature tensors are all processed sequentially by a convolutional layer, a group of residual blocks, and an upsample layer implemented by bilinear interpolation. After the feature tensors from these six paths are gathered, the concatenate layer is applied for the feature combination along the channel axis. Finally, the feature tensors are processed by a 2D convolution layer, two residual blocks, and a 2D convolution layer without ReLU to obtain 32-channel feature tensors with 1/4 resolution.

Optics EXPRESS

At the next stage, for constructing the 4D cost volume, feature tensors of each pixel in the left image and all corresponding candidates in the local disparity range on the epipolar line of the right image are concatenated. The initial 4D cost volume of dimensionality $H \times W \times D \times F$ (*i.e.*, *Height* $\times$ *Width* $\times$ *Disparity* $\times$ *Features*) is built as shown in Fig. 4:

$$\begin{aligned} Cost(:, 1:(W-D_i), D_i - D_{min} + 1, 1:\frac{F}{2}) &= Feature_{left}(:, 1:(W-D_i), :), \\ Cost(:, 1:(W-D_i), D_i - D_{min} + 1, (\frac{F}{2}+1):F) &= Feature_{right}(:, (D_i+1):W, :) \end{aligned}, \tag{13}$$

where $Feature_{left}$ and $Feature_{right}$ represent the feature tensors with 1/4 resolution from two perspectives output by the feature extraction subnetwork, their size ($H \times W \times F/2$) is $240 \times 320 \times 32$ for the 480×640 resolution of the cameras. $[2D_{min}, 2D_{max}]$ is the disparity range of our SPP system. For feature tensors with 1/4 resolution, the initial 4D cost volume is built based on the range $[D_{min}, D_{max}]$. D_i is a candidate disparity in the range $[D_{min}, D_{max}]$. D is the absolute disparity range ($D_{max} - D_{min} + 1$).

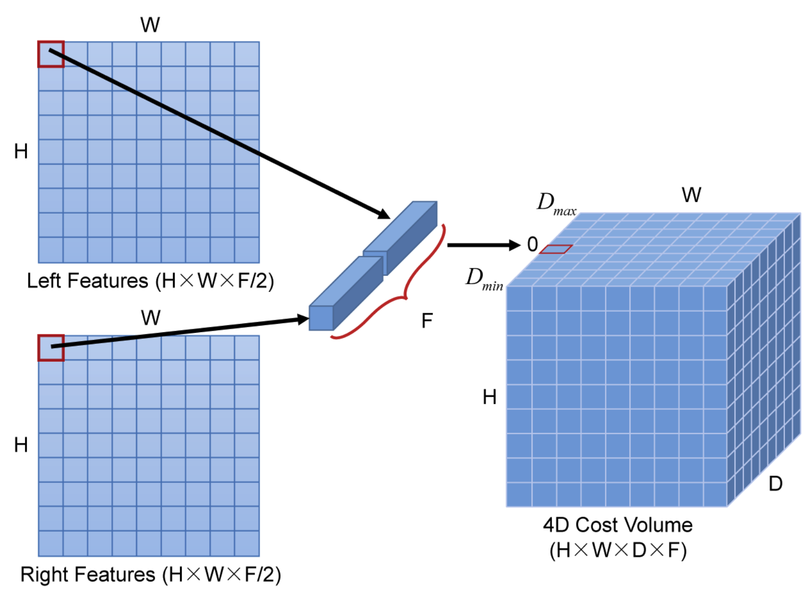

Fig. 4. The schematic diagram of the construction of the 4D cost volume. Based on the disparity range of our SPP system, the initial 4D cost volume is built by combining feature tensors of each pixel in the left image and all corresponding candidates along the epipolar line of the right image.

In cost aggregation, the initial 4D cost volume will be further optimized using 3D convolutional layers. Although some downsample operations have been done during feature extraction, in fact, the 4D cost volume with 1/4 resolution still occupies a lot of GPU memories. Therefore, a lightweight 3D U-net network is proposed to achieve efficient 4D cost aggregation. First of all, three sets of 3D convolutional layers are adopted to realize cost filtering and downsample the 4D cost volume by 1/4. Then, the ConvTranspose3d layer is used to upsample the cost volume, and combined with shortcut operations to achieve residual aggregation. According to the output of the residual operations, three 3D convolutional layers are used to acquire a 4D cost volume with a single-channel feature, and subsequently obtain the final full-resolution 4D cost volume through an upsample layer.

Disparity regression in [49] is introduced to estimate the disparity map based on the final 4D cost volume with a single-channel feature. The probability of each candidate disparity D_i is first calculated using the softmax operation for the predicted cost volume. The predicted disparity map $Disparity(x, y)$ is procured by the weighted sum of the normalized probability for each candidate disparity D_i:

$$Disparity(x, y) = \sum_{D_i=2D_{min}}^{2D_{max}} D_i \times softmax(Cost(x, y, D_i)). \tag{14}$$

The traditional stereo matching network directly calculates the loss between the predicted disparity map and the ground-truth for training. But for the dataset built in our SPP system, the disparity map of the sample data has valid values only in the foreground. Therefore, it is necessary to integrate an additional saliency detection network into our existing network. Currently, the learning-based saliency detection method has been widely investigated with its advantages of high accuracy, high efficiency, and low cost. Among them, fully convolutional network (FCN) is one of the most promising network architectures and has achieved significant results on various well-known datasets [56]. However, given the dataset of SPP that the spatial structure of the tested scenes is relatively simple and the saliency objects have strong speckle texture information, a saliency detection network based on a simple network structure can also achieve good detection results. In order to avoid extracting redundant features, the feature tensors from two perspectives output by the feature extraction subnetwork are directly stacked through a concatenate layer. And then, through a group of residual blocks, a ConvTranspose2d layer, another group of residual blocks, and a convolutional layer, the feature tensors are sequentially filtered and upsampled to obtain a single-channel feature tensor with full resolution. Finally, the sigmoid function is used to achieve the regression of the saliency detection mask $Mask(x, y)$, enabling the prediction of the disparity map without the background:

$$Disparity_{train}(x, y) = Disparity(x, y) \times Mask(x, y). \tag{15}$$

During training, we used *Adam* to minimize the joint loss, thereby updating the weights that parameterize the network. The joint loss consists of a smooth L1 loss for the disparity map and a binary cross-entropy loss for the saliency mask:

$$Loss = Loss_{Mask} + Loss_{Disparity}, \tag{16}$$

$$Loss_{Mask} = -\frac{1}{N}\sum_{n=1}^{N}[Mask_v^c(n)\ln Mask(n) + (1 - Mask_v^c(n))\ln(1 - Mask(n))], \tag{17}$$

$$Loss_{Disparity} = \frac{1}{N}\sum_{n=1}^{N} smooth_{L_1}(D_{sub}(n) - Disparity_{train}(n)), \tag{18}$$

$$smooth_{L_1}(x) = \begin{cases} 0.5x^2, & if\ |x|<1, \\ |x| - 0.5, & otherwise. \end{cases} \tag{19}$$

where $Mask_v^c$ and D_{sub} are the corresponding ground truth of the saliency mask and the disparity map according to Section 2.1.

During testing, the saliency detection mask $Mask(x, y)$ needs to be binarized to distinguish the foreground from the background, and the final disparity map is obtained:

$$Disparity_{final}(x, y) = Disparity(x, y), \quad if\ Mask(x, y)>=0.75. \tag{20}$$

To verify the actual impact of the saliency detection network, the comparison of the 3D reconstruction results without/with the saliency detection network is presented as shown in Fig. 5.

It can be found in Fig. 5 that our measurement results without the saliency detection network have serious mismatches in the background, which will affect the convergence of the network during training and reduce the actual performance of the network. Therefore, the saliency detection network is an additional but necessary module in our approach, implicitly enhancing the matching accuracy for valid pixels.

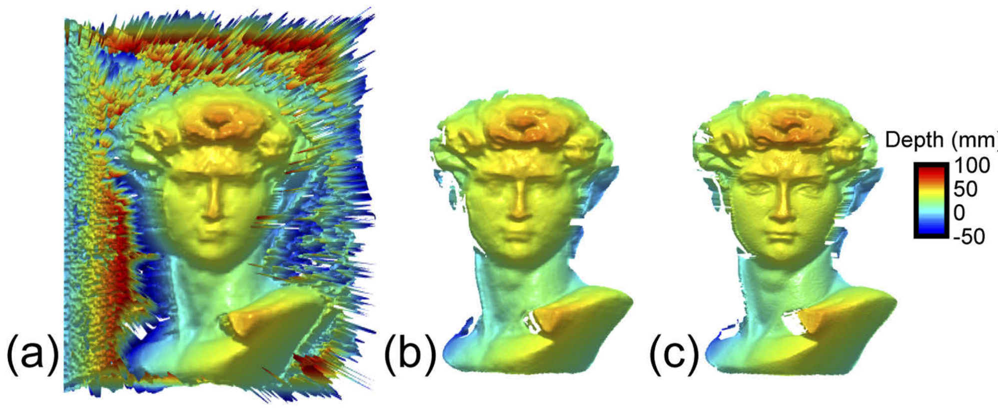

Fig. 5. Comparison of the 3D reconstruction results without/with the saliency detection network. (a) the 3D reconstruction results without the saliency detection network. (b) the 3D reconstruction results with the saliency detection network. (c) the ground truth.

3. Experiments

To verify the actual 3D measurement performance of the proposed method, a common stereo vision-based SPP system with a wide baseline is built as shown in Fig. 1, which consists of two monochrome cameras (Basler acA640-750um with the resolution of 640×480) and a DLP projector (LightCrafter 4500Pro with the resolution of 912×1140). Since the baseline between the stereo cameras is about $270mm$, the disparity constraint of our system should be suitably set to -100 to 59 pixels to measure objects with a depth range of $-100mm$ to $100mm$. The distance between the measurement system and the objects to be tested is about $900mm$. In addition, the projected speckle pattern has been designed and evaluated based on our previous work [31] to obtain the best 3D measurement performance.

In our experiment, we collected the dataset including 1200 different scenes, which are randomly composited 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. During training, to monitor the accuracy of the neural networks for samples that they have never seen, the scenes in these training, verification, and testing datasets are separate from each other. In addition, to achieve high-robustness and high-accuracy stereo matching, the proposed stereo matching network can only process a pair of stereo images at a time during training, which occupies about 23GB graphic memories. The training epoch is set as 200 which takes about 5 days. The proposed network takes 0.95 seconds for disparity prediction.

3.1. Experimental comparison of different methods

A comparative experiment is first carried out to reveal the high performance of the proposed method compared with two traditional methods (ZNCC [57] and SGM_Census [41,42]) and two learning-based methods (Luo's method [46] and BM_DL proposed in our previous work [55]). Measuring the objects with ridged, complex, or discontinuous surfaces is a challenging task for a single-shot SPP. To verify the reliability of these methods for scanning these challenging

Optics EXPRESS

surfaces, two different objects are measured including the David model and the statue of Voltaire. The corresponding 3D reconstruction results obtained by ZNCC, SGM_Census, Luo's method, BM_DL, and our method are shown in Figs. 6(a) and (c).

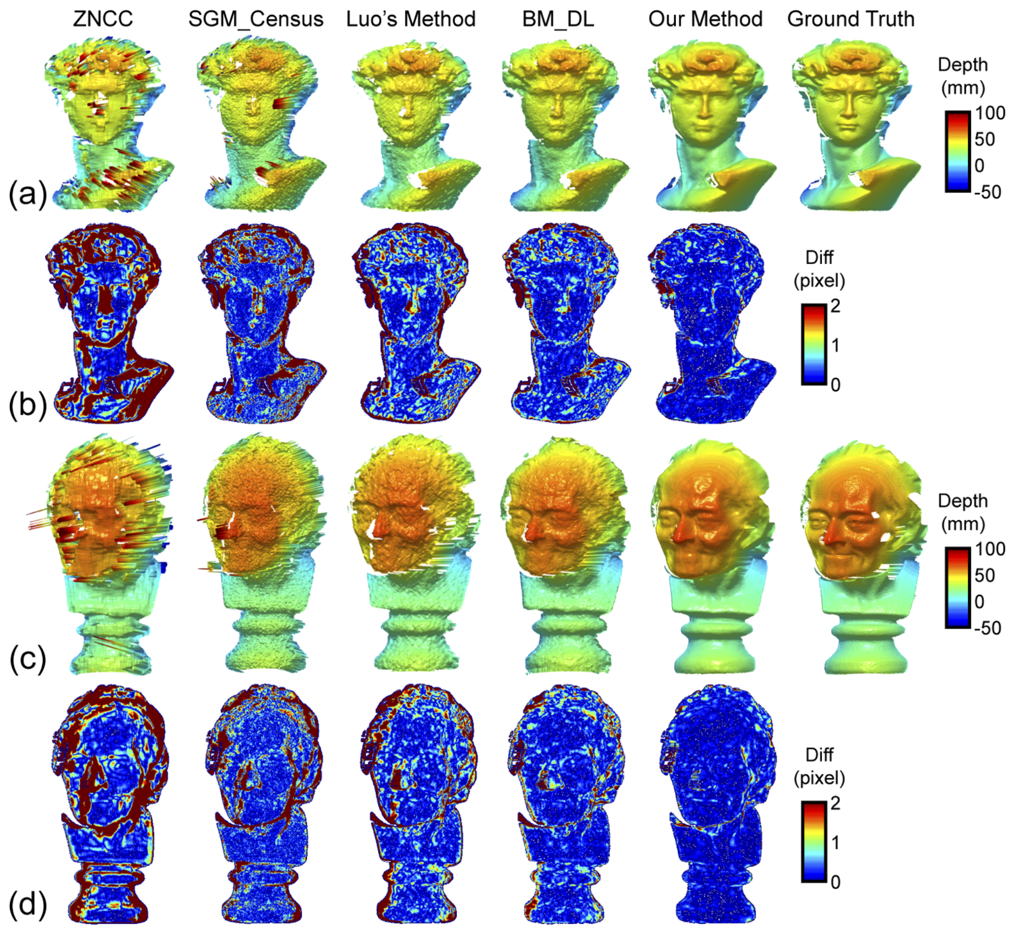

Fig. 6. Comparison of the 3D reconstruction results using different methods. (a) the 3D reconstruction results of the David model, (b) the matching errors of the David model, (c) the 3D reconstruction results of the statue of Voltaire, (d) the matching errors of the statue of Voltaire.

The ZNCC criterion is highly common for practical use, as it is insensitive to the offset and scale changes in the intensity of the local matched block and provides the most accurate and reliable displacement estimations compared with other criteria [57]. In ZNCC, block matching is performed to calculate the matching costs and acquire the integer-pixel disparity maps, which then can be refined to obtain the sub-pixel disparity maps by a five-point quadratic curve fitting model [14]. In order to enhance the matching performance of ZNCC, the block size in block matching is determined as 19×19 after an exhaustive empirical search. However, the fundamental assumption made by block matching is that all the pixels in the matching window have similar disparities. As a consequence, this assumption does not hold at disparity discontinuities, causing the corresponding 3D results with the edge-fattening issue [58,59] in object boundaries and thin structures as shown in Fig. 6.

Optics EXPRESS

Compared with ZNCC, SGM_Census can provide dense 3D measurement results. In SGM_Census, the census transform with the same block size of 19×19 is applied to calculate the initial matching costs, which are then processed to obtain the 3D results using a series of post-processing operations including 1D cost aggregation from 8 paths, Winner-Take-All (WTA), and a quadratic curve fitting [41]. However, SGM_Census avoids mismatching by smoothing the disparity map for achieving reliable stereo matching, at the cost of 3D measurement accuracy as shown in Fig. 6. It can be found that there are some obvious mismatch areas and low-precision 3D measurement results using ZNCC and SGM_Census, which proves that these non-parametric matching methods are so difficult to provide reliable and high-precision matching results on the SPP system with a wide baseline.

Different from these traditional methods, two learning-based methods (Luo's method and BM_DL) are also implemented for comparison. In the two methods, matching cost calculation is implemented using the network. In Luo's method, a pair of block data (centered on the point to be matched in the left image and its all corresponding candidate points in the right image) is inputted into the network at the same time to search the correct candidate point within the pre-defined local disparity range. To realize the high performance of stereo matching, a block matching network based on the Siamese structure is adopted to generate better initial matching costs. Similar to SGM_Census, a series of same post-processing operations are used to obtain the 3D results as shown in Fig. 6. Furthermore, BM_DL proposed in our previous work is an enhanced version of Luo's method. In the block matching network of BM_DL, 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 find in Fig. 6 that BM_DL can output more accurate and dense disparity results compared with SGM_Census and Luo's method. However, the measurement accuracy achieved by BM_DL cannot meet the requirements of high-precision 3D measurement applications. It is important that how to leverage the end-to-end network to achieve more efficient three-dimensional matching is worth investigating.

Obviously, in Fig. 6, the proposed end-to-end stereo matching network yields the highest-quality 3D reconstruction by the single-shot measurement. Compare with the ground truth using the 12-step phase-shifting fringe patterns as shown in Fig. 6, due to the inherent characteristics of local smoothness for stereo matching, there are some local details with slight distortion and blurred surfaces in our 3D reconstruction results. However, it can be found that our method can obtain high-precision 3D results that are closer to the ground truth. It is easy to conclude based on these experimental results that our matching network can achieve 3D measurements with the best performance among several SPP methods.

Besides, compared with the ground truth, the matching errors for different methods are shown in Figs. 6(b) and 6(d) and the corresponding quantitative analysis results can be found in Table 1. To ensure the objectivity of the analysis results, the differences between the disparity results obtained using these methods and the ground truth are used to make an accurate judgment. The number of points is the sum of valid points in the ground truth. The missing ratio means the proportion of points that are valid points in the ground truth but invalid points in these disparity results. For ZNCC, SGM_Census, Luo's method, and BM_DL, the 4-connected image segmentation method is used to process the disparity maps to identify and remove segments with fewer pixels [41]. For our method, the mask generated by the saliency detection subnetwork is exploited to directly remove the invalid pixels in the disparity maps including occlusions and backgrounds. Then the error ratio is easily obtained by counting the number of valid points where their absolute disparity difference between the ground truth and these disparity results are more than 1 pixel. All remaining valid points are regarded as correct points and then further subdivided according to different disparity accuracies including 1 pixel, 0.5 pixels, and 0.2 pixels. It can

Optics EXPRESS

be seen from Table 1 that the missing ratio and the error ratio using our method are lower than 2% and 6%. The correctness ratio achieved by our method is higher than 93%, and most of the pixels have a disparity accuracy of lower than 0.5 pixels. The results illustrated that the matching accuracy using the proposed method is improved by about 50% significantly compared with traditional stereo matching methods. Our method can achieve robust 3D shape measurement with a high correctness ratio and high completeness for objects with complex surfaces and geometric discontinuities.

Table 1. Quantitative analysis results for different methods

Object	Nop[a]	Method	Mmr[b](%)	Emr[c](%)	Cmr[d](%)		
					≤ 1[e]	≤ 0.5[f]	≤ 0.2[g]
David	62337	ZNCC	16.92	34.77	48.31	33.98	16.93
		SGM_Census	10.09	22.18	67.73	49.21	23.73
		Luo's method	7.91	18.26	73.83	52.36	25.05
		BM_DL	3.44	13.24	83.32	64.26	33.18
		Our method	**1.57**	**5.34**	**93.09**	**83.67**	**56.79**
Voltaire	75403	ZNCC	10.91	28.07	61.02	45.92	22.63
		SGM_Census	6.11	18.51	75.38	56.96	28.13
		Luo's method	6.97	14.71	78.32	59.06	39.31
		BM_DL	2.29	9.69	88.02	72.11	39.48
		Our method	**0.68**	**2.84**	**96.48**	**89.48**	**62.64**

[a]Nop = Number of points,
[b]Mmr = Missing matching rate,
[c]Emr = Error matching rate,
[d]Cmr = Correct matching rate,
[e]≤ 1 = Less than 1 pixel,
[f]≤ 0.5 = Less than 0.5 pixels,
[g]≤ 0.2 = Less than 0.2 pixels.

3.2. Precision analysis

Further, to quantitatively evaluate the accuracy of our system using the proposed end-to-end stereo matching network, a ceramic plane and a pair of standard ceramic spheres with a diameter of 50.8mm are measured. Figures 7(a) and 7(b) show the corresponding 3D reconstruction results. And then, based on the obtained 3D reconstruction data, the plane fitting is performed to acquire the ideal plane as the ground truth. The difference between the measured plane and the ideal plane is calculated to obtain the 3D measured errors as shown in Fig. 7(c). The quantitative histograms of the differences are displayed as shown in Fig. 7(f). It can be easily found that the major measured errors are less than 200μm with the RMS of 101.65μm, respectively. Likewise, for the 3D measurement of a pair of standard ceramic spheres as shown in Fig. 7(b), the sphere fitting is used to obtain the actual measurement error as shown in Figs. 7(d) and 7(e). Then, the RMS of the 3D measurement accuracy is about 100μm as shown in Figs. 7(g) and 7(h).

In addition, the precision analysis results for different methods are presented in Table 2. For the ceramic plane, the measurement errors achieved using ZNCC are less than 200μm with the RMS of 103.04μm. The reason for this result is that based on the basic assumption of block matching all pixels in the matching window have similar disparities. However, this assumption does not hold for measuring objects with ridged, complex, or discontinuous faces. For the standard ceramic spheres, ZNCC can only generate coarse 3D measurement results with many matching errors as shown in Fig. 8. It is noted that by the sphere fitting the actual measurement errors are greater than 1mm. After outlier removal, the measurement accuracy has been improved

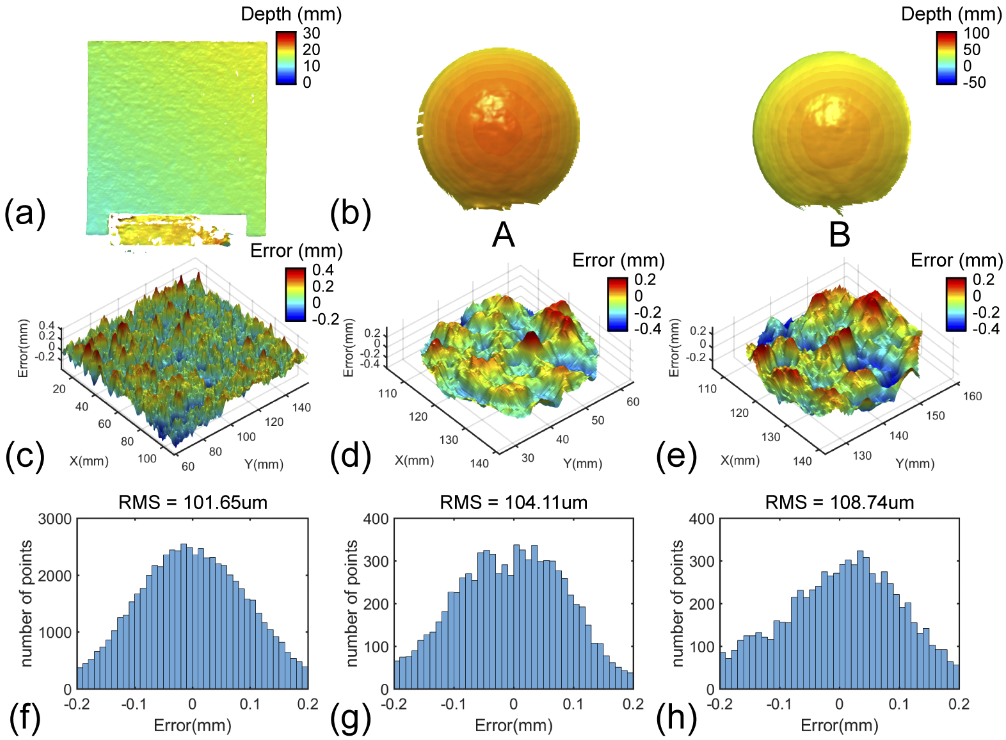

Fig. 7. Precision analysis for measuring a ceramic plane and a pair of standard ceramic spheres using our method. (a) The 3D reconstruction results of a ceramic plane, (b) the 3D reconstruction results of a pair of standard ceramic spheres, (c)-(e) the corresponding distributions of the measured errors of (a)-(b), and (f)-(h) the corresponding quantitative histograms of the measured errors of (a)-(b).

significantly but is still greater than $300\mu m$. And the radius error of the tested ceramic spheres using ZNCC is greater than $1mm$ in Table 2. In contrast, SGM_Census provides measurement results with similar accuracy for measuring planes and spheres. Similarly, Luo's method and BM_DL can also realize robust and more accurate measurements for measuring planes and spheres. However, these methods all use the same post-processing operations to achieve reliable stereo matching by smoothing the disparity map, at the cost of matching accuracy. Unlike these methods, whether the planes or spheres are measured, and whether RMS or radius errors of the spheres are calculated, our method can achieve robust 3D shape measurement with the best accuracy. This result verifies that the proposed method can significantly increase the matching accuracy of SPP and achieve high-precision 3D reconstruction results.

Table 2. Precision analysis results for different methods

Method	Ceramic plane	Ceramic sphere A		Ceramic sphere B	
	RMS	RMS	Radius error	RMS	Radius error
ZNCC	$103.04\mu m$	$340.62\mu m$	$1.31mm$	$367.25\mu m$	$1.02mm$
SGM_Census	$279.39\mu m$	$332.11\mu m$	$244.35\mu m$	$345.76\mu m$	$260.72\mu m$
Luo's method	$240.12\mu m$	$292.65\mu m$	$213.05\mu m$	$274.82\mu m$	$229.35\mu m$
BM_DL	$189.85\mu m$	$213.66\mu m$	$172.85\mu m$	$207.79\mu m$	$166.57\mu m$
Our method	$\mathbf{101.65\mu m}$	$\mathbf{104.11\mu m}$	$\mathbf{117.85\mu m}$	$\mathbf{108.74\mu m}$	$\mathbf{114.13\mu m}$

Optics EXPRESS

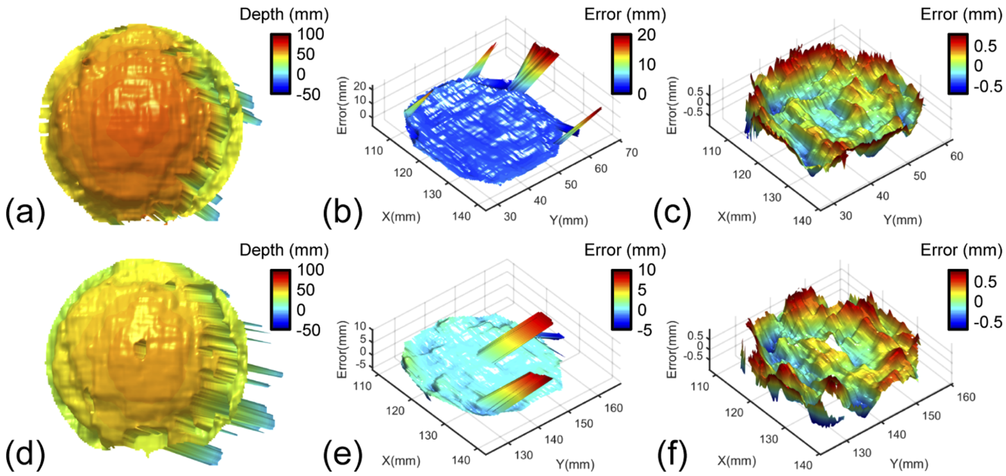

Fig. 8. Precision analysis for measuring a pair of standard ceramic spheres using ZNCC. (a) The 3D reconstruction results of ceramic spheres A, (b) the corresponding distributions of the measured errors of (a), (c) the corresponding distributions of the measured errors of (a) after outlier removal, (d) the 3D reconstruction results of ceramic spheres B, (e) the corresponding distributions of the measured errors of (d), and (f) the corresponding distributions of the measured errors of (d) after outlier removal.

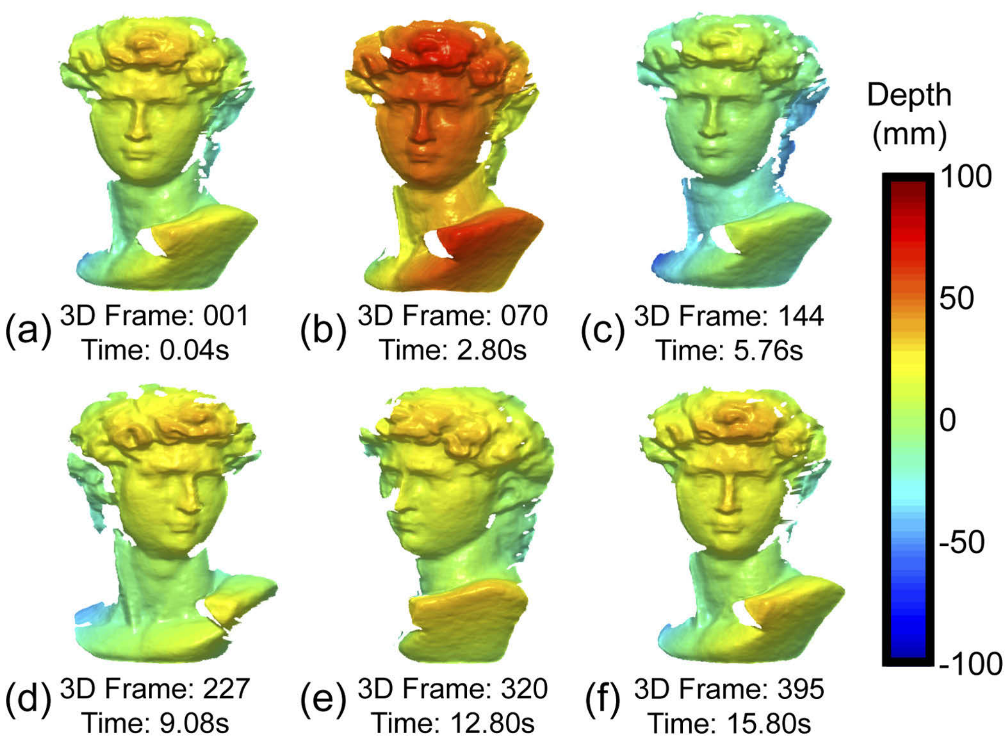

Fig. 9. The 3D reconstruction results for a dynamic scene: a moving David model (Visualization 1). (a)-(c) The David model moves along the Z axis and (d)-(f) the David model rotates around the Y axis.

Optics EXPRESS

3.3. Fast 3D surface imaging

Last of all, our system is applied to record a dynamic scene for fast 3D shape measurement: a moving David model as shown in Fig. 9. In this experiment, the exposure time of cameras is set 39.2ms to capture the speckle images at the speed of 25Hz for achieving 3D reconstruction at 25fps. Figure 9 shows the color-coded 3D reconstruction results at different time points. During the whole dynamic measurement, the David model first moves forward along the Z axis, and arrives at the boundary of the predefined measurement space at 2.8 seconds. Then, the David model moves in reverse along the Z axis to another boundary of the predefined measurement space at 5.76 seconds. Furthermore, the David model returns to the initial position and starts to rotate around the Y axis. Finally, it is back to the origin position again in 15.8 seconds. The whole 3D measurement results can refer to Visualization 1. In the whole measuring procedures, the 3D surfaces of the David model are correctly and high-quality reconstructed, verifying the reliability of the proposed method to perform the absolute 3D shape measurement with high completeness at high speed.

4. Conclusion

In summary, we proposed a single-shot 3D shape measurement method using an end-to-end stereo matching network based on a common stereo vision-based SPP system. To efficiently train the stereo matching network, a high-quality SPP dataset is first built by combining phase-shifting profilometry (PSP) and temporal phase unwrapping techniques in FPP. High-precision absolute phase maps obtained using FPP are used to generate accurate and dense disparity maps with high completeness as the ground truth of the dataset by phase matching. For the architecture of the network, the proposed network first leverages a multi-scale residual subnetwork to synchronously extract compact feature tensors with 1/4 resolution from speckle images for constructing the 4D cost volume. Although some downsample operations have been done during feature extraction, in fact, the 4D cost volume with 1/4 resolution still occupies a lot of GPU memories. Therefore, a lightweight 3D U-net network is proposed to implemented efficient 4D cost aggregation for achieving higher matching performance. Considering that the disparity maps (as the ground truth) in the SPP dataset has valid values only in the foreground, a simple and fast saliency detection network is proposed and integrated into our network to avoid enhancing the invalid pixels in the disparity maps including occlusions and backgrounds, thereby implicitly enhancing the matching accuracy for valid pixels. The experimental comparison of different methods illustrated that compared with traditional methods our method can achieve robust 3D shape measurement with a high correctness ratio and high completeness for objects with complex surfaces. Besides, the quantitative analysis results proved again that the matching accuracy using the proposed method is improved by about 50% significantly compared with traditional stereo matching methods. The experiment results of the precision analysis demonstrated that the proposed method can achieve absolute 3D shape measurement with an accuracy of about 100μm through only a single speckle pattern. The dynamic measurement experiment has verified the success of the proposed method in its ability to effectively achieve fast and accurate 3D shape measurements with high completeness for complex scenes at 25fps.

Finally, there are several aspects that need to be further improved in the proposed method. First, since there are many costly 3D convolutions for cost aggregation in the proposed network, the initial cost volume is 1/4 downsampled in advance, which undoubtedly reduces the accuracy of stereo matching significantly. Therefore, how to achieve more efficient cost aggregation is still a problem to be solved. Second, it is easy to understand that projecting multiple speckle images will improve the accuracy of 3D measurement, because more constraints can be exploited to completely guarantee the global uniqueness of the measured scenes. How to improve the measurement accuracy of the stereo matching network by inputting multiple speckle images at the same time is another interesting direction for further investigation. Third, the proposed network takes 0.95 seconds for disparity prediction that is slower compared with most of the existing algorithms running on GPU. How to achieve fast stereo matching should be considered. It can be found that cost aggregation in the proposed network take accounts for most of the total run time. Similarly, the cost aggregation sub-network should be further optimized to improve the accuracy of stereo matching and reduce the run time. At last, different from traditional non-learning methods, it is noted that the generalization ability of learning methods needs to be further researched and discussed for measuring different objects with complex reflection characteristics or high reflectivity, enabling more reliable 3D shape measurement. Based on the above analysis, we will explore more other methods to design a single-shot SPP system with higher performance.

Funding. National Defense Science and Technology Foundation of China (2019-JCJQ-JJ-381); National Key Research and Development Program of China (2017YFF0106403); Leading Technology of Jiangsu Basic Research Plan (BK20192003); "333 Engineering" Research Project of Jiangsu Province (BRA2016407); Open Research Fund of Jiangsu Key Laboratory of Spectral Imaging and Intelligence Sense (3091801410411); Fundamental Research Funds for the Central Universities (30919011222, 30920032101); National Natural Science Foundation of China (61705105, 61722506, 62005121, 62075096).

Disclosures. The authors declare no conflicts of interest.

References

1. S. S. Gorthi and P. Rastogi, "Fringe projection techniques: whither we are?" Opt. Laser Eng. **48**(2), 133–140 (2010).
2. S. Feng, L. Zhang, C. Zuo, T. Tao, Q. Chen, and G. Gu, "High dynamic range 3d measurements with fringe projection profilometry: a review," Meas. Sci. Technol. **29**(12), 122001 (2018).
3. Z. Zhang, "Review of single-shot 3d shape measurement by phase calculation-based fringe projection techniques," Opt. Laser Eng. **50**(8), 1097–1106 (2012).
4. W. Yin, S. Feng, T. Tao, L. Huang, S. Zhang, Q. Chen, and C. Zuo, "Calibration method for panoramic 3d shape measurement with plane mirrors," Opt. Express **27**(25), 36538–36550 (2019).
5. Q. Zhang and X. Su, "High-speed optical measurement for the drumhead vibration," Opt. Express **13**(8), 3110–3116 (2005).
6. Z. Zhang, S. Huang, S. Meng, F. Gao, and X. Jiang, "A simple, flexible and automatic 3d calibration method for a phase calculation-based fringe projection imaging system," Opt. Express **21**(10), 12218–12227 (2013).
7. J. Salvi, J. Pages, and J. Batlle, "Pattern codification strategies in structured light systems," Pattern Recognition **37**(4), 827–849 (2004).
8. S. Zhang, "High-speed 3d shape measurement with structured light methods: A review," Opt. Laser Eng. **106**, 119–131 (2018).
9. C. Zuo, T. Tao, S. Feng, L. Huang, A. Asundi, and Q. Chen, "Micro fourier transform profilometry (μftp): 3d shape measurement at 10, 000 frames per second," Opt. Laser Eng. **102**, 70–91 (2018).
10. S. Zhang, "Absolute phase retrieval methods for digital fringe projection profilometry: A review," Opt. Laser Eng. **107**, 28–37 (2018).
11. W. Yin, C. Zuo, S. Feng, T. Tao, Y. Hu, L. Huang, J. Ma, and Q. Chen, "High-speed three-dimensional shape measurement using geometry-constraint-based number-theoretical phase unwrapping," Opt. Laser Eng. **115**, 21–31 (2019).
12. M. Schaffer, M. Grosse, B. Harendt, and R. Kowarschik, "High-speed three-dimensional shape measurements of objects with laser speckles and acousto-optical deflection," Opt. Lett. **36**(16), 3097–3099 (2011).
13. M. Schaffer, M. Grosse, and R. Kowarschik, "High-speed pattern projection for three-dimensional shape measurement using laser speckles," Appl. Opt. **49**(18), 3622–3629 (2010).
14. P. Zhou, J. Zhu, and H. Jing, "Optical 3-d surface reconstruction with color binary speckle pattern encoding," Opt. Express **26**(3), 3452–3465 (2018).
15. X. Su and W. Chen, "Fourier transform profilometry: a review," Opt. Laser Eng. **35**(5), 263–284 (2001).
16. Q. Kemao, "Two-dimensional windowed fourier transform for fringe pattern analysis: principles, applications and implementations," Opt. Laser Eng. **45**(2), 304–317 (2007).
17. S. Feng, Q. Chen, G. Gu, T. Tao, L. Zhang, Y. Hu, W. Yin, and C. Zuo, "Fringe pattern analysis using deep learning," Adv. Photonics **1**(2), 025001 (2019).
18. C. Zuo, S. Feng, L. Huang, T. Tao, W. Yin, and Q. Chen, "Phase shifting algorithms for fringe projection profilometry: A review," Opt. Laser Eng. **109**, 23–59 (2018).
19. X. Su and W. Chen, "Reliability-guided phase unwrapping algorithm: a review," Opt. Laser Eng. **42**(3), 245–261 (2004).
20. M. Zhao, L. Huang, Q. Zhang, X. Su, A. Asundi, and Q. Kemao, "Quality-guided phase unwrapping technique: comparison of quality maps and guiding strategies," Appl. Opt. **50**(33), 6214–6224 (2011).
21. Y. Wang and S. Zhang, "Novel phase-coding method for absolute phase retrieval," Opt. Lett. **37**(11), 2067–2069 (2012).

22. C. Zuo, L. Huang, M. Zhang, Q. Chen, and A. Asundi, "Temporal phase unwrapping algorithms for fringe projection profilometry: A comparative review," Opt. Laser Eng. **85**, 84–103 (2016).
23. K. Zhong, Z. Li, Y. Shi, C. Wang, and Y. Lei, "Fast phase measurement profilometry for arbitrary shape objects without phase unwrapping," Opt. Laser Eng. **51**(11), 1213–1222 (2013).
24. X. Liu, Y. Yang, Q. Tang, Z. Cai, X. Peng, M. Liu, and Q. Li, "A method for fast 3d fringe projection measurement without phase unwrapping," in *Sixth International Conference on Optical and Photonic Engineering (icOPEN 2018)*, vol. 10827 (International Society for Optics and Photonics, 2018), p. 1082713.
25. W. Yin, Q. Chen, S. Feng, T. Tao, L. Huang, M. Trusiak, A. Asundi, and C. Zuo, "Temporal phase unwrapping using deep learning," Sci. Rep. **9**(1), 20175 (2019).
26. K. Liu, Y. Wang, D. L. Lau, Q. Hao, and L. G. Hassebrook, "Dual-frequency pattern scheme for high-speed 3-d shape measurement," Opt. Express **18**(5), 5229–5244 (2010).
27. C. Zuo, Q. Chen, G. Gu, S. Feng, and F. Feng, "High-speed three-dimensional profilometry for multiple objects with complex shapes," Opt. Express **20**(17), 19493–19510 (2012).
28. C. Zuo, Q. Chen, G. Gu, S. Feng, F. Feng, R. Li, and G. Shen, "High-speed three-dimensional shape measurement for dynamic scenes using bi-frequency tripolar pulse-width-modulation fringe projection," Opt. Laser Eng. **51**(8), 953–960 (2013).
29. X. Su and Q. Zhang, "Dynamic 3-d shape measurement method: a review," Opt. Laser Eng. **48**(2), 191–204 (2010).
30. S. Feng, C. Zuo, T. Tao, Y. Hu, M. Zhang, Q. Chen, and G. Gu, "Robust dynamic 3-d measurements with motion-compensated phase-shifting profilometry," Opt. Laser Eng. **103**, 127–138 (2018).
31. W. Yin, S. Feng, T. Tao, L. Huang, M. Trusiak, Q. Chen, and C. Zuo, "High-speed 3d shape measurement using the optimized composite fringe patterns and stereo-assisted structured light system," Opt. Express **27**(3), 2411–2431 (2019).
32. B. Pan, Z. Lu, and H. Xie, "Mean intensity gradient: an effective global parameter for quality assessment of the speckle patterns used in digital image correlation," Opt. Laser Eng. **48**(4), 469–477 (2010).
33. Z. Chen, X. Shao, X. Xu, and X. He, "Optimized digital speckle patterns for digital image correlation by consideration of both accuracy and efficiency," Appl. Opt. **57**(4), 884–893 (2018).
34. M. Ito and A. Ishii, "A three-level checkerboard pattern (tcp) projection method for curved surface measurement," Pattern Recognit. **28**(1), 27–40 (1995).
35. M. Maruyama and S. Abe, "Range sensing by projecting multiple slits with random cuts," IEEE Trans. Pattern Anal. Machine Intell. **15**(6), 647–651 (1993).
36. K. L. Boyer and A. C. Kak, "Color-encoded structured light for rapid active ranging," *IEEE Transactions on Pattern Analysis Mach. Intell.* pp. 14–28 (1987).
37. L. Zhang, B. Curless, and S. M. Seitz, "Rapid shape acquisition using color structured light and multi-pass dynamic programming," in *First International Symposium on 3D Data Processing Visualization and Transmission*, (IEEE, 2002), pp. 24–36.
38. J. Pagès, J. Salvi, C. Collewet, and J. Forest, "Optimised de bruijn patterns for one-shot shape acquisition," Image Vis. Comput. **23**(8), 707–720 (2005).
39. H. Morita, K. Yajima, and S. Sakata, "Reconstruction of surfaces of 3-d objects by m-array pattern projection method," in 1988 IEEE Conference on International Conference on Computer Vision, (IEEE, 1988), pp. 468–473.
40. S. Heist, P. Dietrich, M. Landmann, P. Kühmstedt, G. Notni, and A. Tünnermann, "Gobo projection for 3d measurements at highest frame rates: a performance analysis," Light: Sci. Appl. **7**(1), 71 (2018).
41. H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," IEEE Trans. Pattern Anal. Mach. Intell. **30**(2), 328–341 (2008).
42. H. Hirschmuller and D. Scharstein, "Evaluation of stereo matching costs on images with radiometric differences," IEEE Trans. Pattern Anal. Mach. Intell. **31**(9), 1582–1599 (2009).
43. F. Gu, Z. Song, and Z. Zhao, "Single-shot structured light sensor for 3d dense and dynamic reconstruction," Sensors **20**(4), 1094 (2020).
44. A. Geiger, M. Roser, and R. Urtasun, "Efficient large-scale stereo matching," in *Asian conference on computer vision*, (Springer, 2010), pp. 25–38.
45. J. Zbontar and Y. LeCun, "Computing the stereo matching cost with a convolutional neural network," in *2015 IEEE Conference on Computer Vision and Pattern Recognition*, (IEEE, 2015), pp. 1592–1599.
46. W. Luo, A. G. Schwing, and R. Urtasun, "Efficient deep learning for stereo matching," in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, (IEEE, 2016), pp. 5695–5703.
47. J. Pang, W. Sun, J. S. Ren, C. Yang, and Q. Yan, "Cascade residual learning: A two-stage convolutional neural network for stereo matching," in *2017 IEEE Conference on International Conference on Computer Vision Workshops*, (IEEE, 2017), pp. 887–895.
48. N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, (IEEE, 2016), pp. 4040–4048.
49. A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, "End-to-end learning of geometry and context for deep stereo regression," in *2017 IEEE Conference on International Conference on Computer Vision*, (IEEE, 2017), pp. 66–75.

Optics EXPRESS

50. S. Khamis, S. Fanello, C. Rhemann, A. Kowdle, J. Valentin, and S. Izadi, "Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction," in *2018 IEEE Conference on European Conference on Computer Vision (ECCV)*, (IEEE, 2018), pp. 573–590.
51. J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *2018 IEEE Conference on Computer Vision and Pattern Recognition*, (IEEE, 2018), pp. 5410–5418.
52. F. Zhang, V. Prisacariu, R. Yang, and P. H. Torr, "Ga-net: Guided aggregation net for end-to-end stereo matching," in *2019 IEEE Conference on Computer Vision and Pattern Recognition*, (IEEE, 2019), pp. 185–194.
53. A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, (IEEE, 2012), pp. 3354–3361.
54. R. Hartley and A. Zisserman, *Multiple view geometry in computer vision* (Cambridge University, 2003).
55. W. Yin, J. Zhong, S. Feng, T. Tao, J. Han, L. Huang, Q. Chen, and C. Zuo, "Composite deep learning framework for absolute 3d shape measurement based on single fringe phase retrieval and speckle correlation," JPhysPhotonics **2**, 045009 (2020).
56. A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li, "Salient object detection: A survey," Comp. Visual Media **5**(2), 117–150 (2019).
57. B. Pan, H. Xie, and Z. Wang, "Equivalence of digital image correlation criteria for pattern matching," Appl. Opt. **49**(28), 5501–5509 (2010).
58. D. Min, J. Lu, and M. N. Do, "A revisit to cost aggregation in stereo matching: How far can we reduce its computational redundancy?" in *2011 International Conference on Computer Vision*, (IEEE, 2011), pp. 1567–1574.
59. D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," Int. J. Comput. Vis. **47**(1/3), 7–42 (2002).