Research Article | Vol. 33, No. 18 / 8 Sep 2025 / Optics Express 37532
Optics EXPRESS

Multichannel autostereoscopic measurement system for micro-structured surfaces based on multi-scale depth fusion

YONGQIANG YANG,[1] CHI FAI CHEUNG,[1,3] DA LI,[1,2,4] AND SANSHAN GAO[1]

[1]State Key Laboratory of Ultra-Precision Machining Technology, Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR, China
[2]Institute of Morden Optics, Nankai University, Tianjin 30071, China
[3]Benny.Cheung@polyu.edu.hk
[4]da.li@nankai.edu.cn

Abstract: Autostereoscopic technology has been acknowledged to realize precise metrology of micro-structured surfaces. While autostereoscopic performs well, the reliance on a single light field modality constrains the performance of the system. To address this issue, a multichannel autostereoscopic measurement (MAM) system has been developed to provide richer data for depth estimation. The whole system includes a 3D optical channel to capture elemental images (EIs) from various viewpoints, in conjunction with a 2D channel to obtain high-resolution (HR) images. This system employs data fusion techniques to compensate for data deficiencies and enhance accuracy. In the 2D channel, a deep learning network called UniDepth is used to estimate the 3D geometry of objects based on HR images. Combining multi-scale depth information from different optical channels via the pyramid representation allows for more precise 3D reconstruction. The experimental results demonstrate that the proposed multichannel measurement system improves the quality and robustness of 3D reconstruction.

1. Introduction

Micro-structured surfaces hold significant potential for diverse applications, including space, optoelectronics [1], biomedical fields [2], and 3D sensing [3]. Advances in ultra-precision manufacturing have spurred interest in measuring these complex surfaces. To meet demands of accuracy, various measurement methods have been developed, categorized into contact and non-contact techniques. Among contact methods, the coordinate measuring machine (CMM), a high-precision measurement tool, is widely used in manufacturing and precision engineering. This equipment operates by moving a probe to designated points on the object being measured. It captures the 3D coordinates of those points and subsequently reconstructs the geometry of the object using the data. Contact measurement techniques are renowned for their high accuracy and reliability [4]. However, their application is hindered by several limitations. Probe size restricts access to intricate features, and physical contact may damage the surface. Additionally, these methods exhibit low efficiency, which impedes rapid data acquisition in large-scale or time-sensitive measurement tasks.

To address these limitations, non-contact methods leverage optical [5], acoustic [6], or other physical principles to acquire data via sensors. These methods avoid the potential for damage or deformation that arises from direct physical interaction with the object and facilitate rapid and repeatable measurements. Structured light, categorized as a non-contact measurement technique, has been developed into a highly sophisticated and mature methodology for 3D measurement. Yin et al. [7] proposed a high-speed 3D shape measurement technology by using composite fringe patterns and structured light. However, the size of the stripe restricts the application of

#567001 https://doi.org/10.1364/OE.567001
Journal © 2025 Received 6 May 2025; revised 21 Jul 2025; accepted 10 Aug 2025; published 26 Aug 2025

Optics EXPRESS

structured light in the measurement. Unlike structured light, which relies on pattern decoding, laser triangulation employs the principles of optical reflection and the concept of similar triangles within the spatial propagation of light. Li et al. [8] proposed a color error compensation to improve the accuracy of the laser triangulation. While this method offers the benefits of high precision and rapid measurement, it depends on the reflection of a discrete point projected onto the surface of objects.

The autostereoscopic technology, categorized as a non-contact method, acquires multi-perspective information in a single snapshot. This enables efficient reconstruction of 3D surface profile. Li et al. [9,10] designed a three-dimensional system and pioneered the integration of autostereoscopic technology. This system enables on-machine three-dimensional metrology. It leverages a micro-lens array (MLA) within the optical channel to disperse light from measured objects into elemental images (EIs). These EIs are 2D images captured by image units in a single optical channel. Employing these images enables the extraction of depth information through processes such as recording, refocusing, and depth estimation. The depth estimation in such systems employs various methods and techniques such as epipolar plane image (EPI) [11], light field refocusing [12], cost volume [13], deep learning [14]. However, these methods typically rely on data from a single optical channel as their input, which imposes certain constraints on their performance. Single channel data is inherently limited by factors such as trade-off between spatial and angular resolution, reduced accuracy due to a small baseline and limited field of view (FOV). All these limitations directly compromise the accuracy of depth estimation, resulting in depth maps that exhibit blurring, discontinuities or unnatural transitions at object boundaries. To address these shortcomings, incorporating broader data is essential. This compensates for the deficiencies in depth caused by relying on a single light field modality and enhances the overall robustness and precision of the system.

This paper presents a multichannel autostereoscopic measurement (MAM) system for micro-structured surfaces, based on a common-path optical architecture. Unlike traditional autostereoscopic measurement systems, which solely obtain depth information from a single 3D optical channel, the proposed system enables simultaneous acquisition of data from both 2D and 3D channels, thereby mitigating the limitations of isolated channel dependency. Both modalities of data are applicable for depth estimation. Specifically, a deep learning model named UniDepth is developed to generate depth information based on the images from 2D channel. Departing from conventional approaches, this model generates HR depth maps preserving more details. The shape from focus (SFF) technique [15] is employed to derive precise depth cues based on EIs captured from 3D channel. Depth information from different channels is inherently complementary, laying a robust foundation for data fusion. The proposed system employs a pyramid representation to fuse multi-scale depth information from these diverse channels. This process mitigates the limitations of individual modalities while improving robustness to noises and environmental variability. For the qualitive analysis, multichannel system reduces discontinuities or unnatural transitions in the depth maps. Compared with the traditional single-optical channel autostereoscopic systems, the proposed system is greatly improved in accuracy and standard deviation. In Section 2, the multichannel 3d autostereoscopic measurement principle is briefly described, and the proposed depth estimation method and depth fusion strategy are illustrated in Section3. The setup for the MAM system as well as experiment results are detailed in Section 4, and the conclusion is provided in Section 5.

2. Multichannel 3D autostereoscopic measurement principle

Autostereoscopic serves as an advanced measurement technology, capable of capturing 3D information in a single shot. The measurement process comprises three 3 key steps: recording, 3D representation, and depth output. In the recording process illustrated in Fig. 1, an MLA positioned before the charge-coupled device (CCD) sensor disperses light from measured objects

Optics EXPRESS

into EIs. This configuration simultaneously captures both spatial information and directional ray information, effectively recording the 4D light field. The distance between corresponding points across EIs is equal to disparities in the EIs. The disparities are directly related to the depth of the points on the measured surfaces. Given that the entire measurement system is stabilized, the disparities observed within the EIs are associated with depth information. These EIs, along with their corresponding disparities, are subsequently utilized in the depth output process. The 3D representation process exhibits symmetry with the recording process due to the reversibility of optical rays. Leveraging the disparity data embedded with EIs, depth is calculated employing the distance relationship between MLA, image sensor and focused points in EIs. By establishing a correlation between focused points and depth, the refocused images and 3D reconstruction surface can be obtained. Refocusing is accomplished by mapping all the points in EIs to a specific focal stack. After repeating this process, a focal stack is obtained. The depth output process involves calculating precise depth information for the focused regions within each layer of the stack.

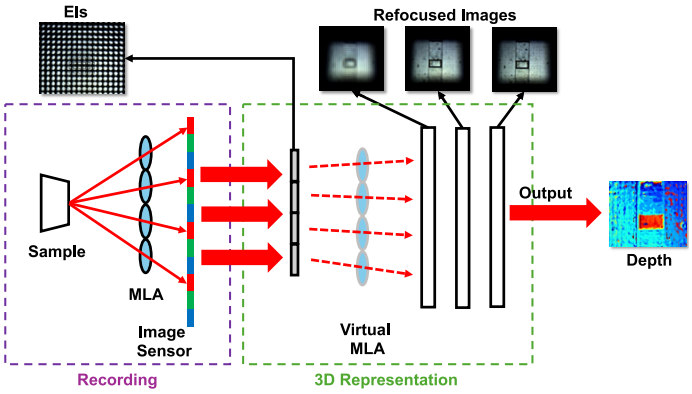

Fig. 1. Autostereoscopic 3D measurement process.

A key limitation of traditional autostereoscopic systems is their reliance on the single optical channel data as input. Single channel data is inherently constrained by factors such as a trade-off between spatial and angular resolution, a small baseline, and a limited FOV. These factors directly affect the accuracy of depth estimation, yielding depth maps that manifest discontinuities and blurring. To address the negative impacts of this dependency on a single channel, a MAM system, as illustrated in Fig. 2, has been designed. The proposed system utilizes the design of the sharing optical path for an autostereoscopic measurement system and a complementary 2D imaging system. The sharing optical path architecture comprises an imaging system, a co-axial illumination and a beam splitter, engineered to ensure optical alignment and efficiency. This configuration enables synchronized multichannel input capture within a unified optical pathway, enhances system stability, reduces environmental inference, and mitigates optical aberrations. In this optical design, beam splitting inherently degrades the signal-to-noise ratio (SNR), adversely affecting reconstruction quality. To address these issues, the beam splitter which has a high beam splitting ratio is employed to optimize light distribution. A high-brightness LED light source is also used to enhance light intensity, thereby improving the SNR. After image acquisition, a Gaussian filter algorithm is applied to mitigate noise and improve image quality. The MAM system enables integrating two optical routes with distinct data modalities. The EIs provide rich angular information critical for depth reconstruction, while the HR 2D images contribute high spatial resolution, enabling precise outputs.

This multichannel design not only improves the quality of depth maps but also enhances the adaptability of the system. Consequently, the MAM system represents a significant advancement

Optics EXPRESS

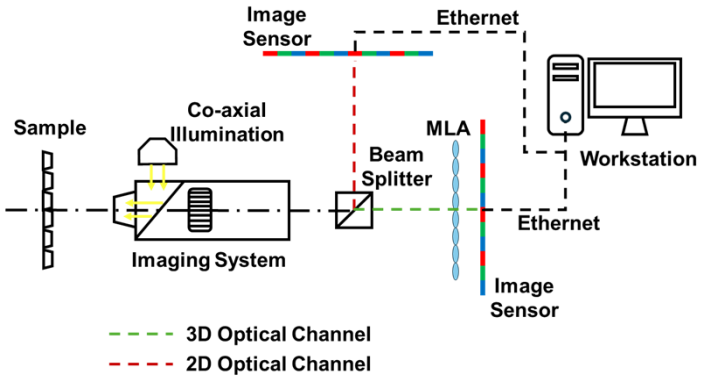

Fig. 2. The design of the MAM system.

in autostereoscopic technology, offering a versatile and effective solution for applications requiring accurate depth perception and detailed imaging.

3. Multichannel autostereoscopic approach based on multi-scale depth fusion

The proposed system is designed to acquire data from multiple channels simultaneously. To fully exploit this multichannel data, advanced methods for depth estimation and depth fusion have been developed at the data processing level. Theoretically, the depth fusion could be indeed enhanced using depth fusion networks, especially the models [16–19] improve fusion accuracy by leveraging multi-level aggregation and multi-scale feature extraction. However, the primary limitation to adopting multiscale depth fusion networks is the scarcity of annotated training data. This limitation poses significant challenges to the effective training of multiscale depth fusion network. To address the current issue, as shown in Fig. 3, a multichannel fusion framework is proposed.

Initially, data such as EIs and HR images are obtained from different channels. Multiple depth estimation techniques such as shape from focus and deep learning are applied to this data, yielding distinct depth maps that reflect complementary perspectives of the measured object. The all-in-focus image generated through refocusing process, serves as the reference image for its texture and intensity. Subsequently, an affine transformation aligns these depth maps and the reference into a unified coordinate system. The depth fusion process employs a sliding window approach. This window, capable of horizontal and vertical slide across an image, selects a region of interest (ROI) in each position. Within each ROI, the pyramid representation decomposes data into multi-scale high frequency (HF) and low frequency (LF) information. During the fusion process, the sliding windows concurrently slide across the depth maps and the reference. In each slide, the HF and LF information generated from two depth maps is compared with that of the reference image. The information exhibiting the highest similarity is preserved, while information that deviates significantly is discarded. This comparison allows for the extraction of optimized HF and LF information, facilitating the creation of a fused pyramid. The comparison is iteratively performed until the sliding is completed. After performing the inverse conversion for the fused pyramid, the fused depth is obtained. For refinement and optimization, a guided filter is employed to smooth the coarse depth and reduce noises.

Optics EXPRESS

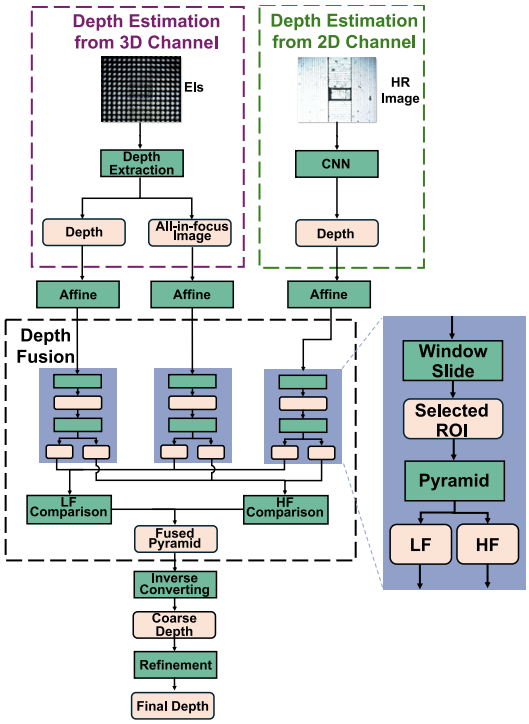

Fig. 3. Framework of the multichannel autostereoscopic approach. The approach receives EIs and HR images as input and enhances the accuracy of measurement results by depth fusion.

3.1. Framework details

3.1.1. Depth estimation from 3D optical channel

Following the acquisition of EIs from the MAM system, the focus shifts to depth estimation based on this data. Shape from focus is a passive depth estimation technique based on the study of focus-defocus cues in a series of images taken at different focal distance. This method has emerged as a powerful approach due to its ability to exploit the rich spatial and angular information captured by CCD sensors. Previous research [15,20,21] has proved that SFF has a good performance in 3D reconstruction and outperforms some well-known techniques. The fundamental principle asserts that the clarity or sharpness of a detected point in an image is maximized when the focal plane of the image system aligns with the depth of that point in the 3D scene. Through systematically assessment of focus quality within a series focal stack, shape from focus infers the depth of each point by determining the focal setting at which its sharpness is maximized. This method, initially designed for conventional imaging systems requiring physical lens adjustment, has been adopted for autostereoscopic systems. This adaptation allows digital refocusing, facilitating the creation of a focal stack from a single shot, improving efficiency and applicability.

In the MAM system, a focal stack is computationally synthesized from the EIs via an MLA. The 4D autostereoscopic data $L(u, v, x, y)$ is refocused to generate images $I_\alpha(x, y)$ at different focal depths. The refocusing process employs the shape from focus algorithm is formulated as:

$$I_\alpha(x, y) = \frac{1}{M \cdot N} \sum_u \sum_v L(u, v, x + u(1 - \frac{1}{\alpha}), y + v(1 - \frac{1}{\alpha})). \tag{1}$$

Optics EXPRESS

where I_α represents focal stack, M and N represent the row and column of an MLA, α is a focal parameter, (u, v) denotes angular coordinates, (x, y) denotes spatial coordinates. Since the shifted coordinates may be non-integer, bilinear interpolation is used to determine the pixel values. Each image in the focal stack requires brightness normalization to mitigate variations in light intensity between different view angles. To quantify the sharpness of local regions for each image in the focal stack, the Sobel operator is applied in the focus measure computation. The Sobel operator is formulated as:

$$K_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, K_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}. \tag{2}$$

For a pixel at position (x, y) in an image $I_\alpha(x, y)$, the gradient magnitude is determined as:

$$F(x, y, \alpha) = \sqrt{(conv(I_\alpha(x, y), K_x))^2 + (conv(I_\alpha(x, y), K_y))^2}. \tag{3}$$

Here, $conv(\cdot)$ denotes the convolution operation, $F(x, y, \alpha)$ represents the focus measure at (x, y) for focal setting α. Thes computations produce a focus curve for each pixel within the focal stack, with peaks indicating optimal focus. For each pixel (x, y), determine the α value yielding the maximum focus measure:

$$\alpha_{\max}(x, y) = \arg\max_{\alpha} F(x, y, \alpha). \tag{4}$$

Shape from Focus is a robust depth estimation technique that capitalizes focus quality variations to reconstruct 3D scene geometry. In autostereoscopic systems, it transforms EIs into a focal stack via refocusing, analyzes sharpness with focus measures, and maps peak focus positions to depths. The strengths of using shape from focus based on Sobel operator lies in edge sensitivity and efficiency. In the MAM system, it contributes to precise depth estimation.

3.1.2. Depth estimation from 2D optical channel

With deep learning emerging as a pivotal technology, it boots the development of measurement systems. Previous advancements in the field of deep learning have demonstrated the feasibility of recovering depth from a single image [22–24]. In the context of the 2D optical channel, the deep learning model serves as an efficient tool, exhibiting strong performance in extracting 3D geometry from 2D images. As shown in Fig. 4, UniDepth model is specifically designed to capture the profile of the micro-structured surfaces.

The UniDepth model includes an encoder and a decoder. It realizes the feature extraction and depth reconstruction. Through the refinement, then final depth is directly used for depth fusion. This network extracts features from the HR images and reconstructs features of depth. The feature extraction is realized by an encoder network that comprises a sequence of double convolution layers, attention mechanisms and down-sampling operations. In the double convolution layer, two consecutive convolution operations are performed, with initial feature channels expanding progressively. The first convolution extracts low-level features such as edges and textures, while the second convolution refines these into high-order representations such as shapes and structures. Each convolution is followed by batch normalization (BN) to stabilize the training process and a ReLU (Rectified Linear Unit) activation function to ensure robust gradient flow. After each double convolution block, the number of feature channels double, enabling the network to capture increasingly abstract features as the spatial resolution decreases.

Following the double convolution, the encoder integrates attention mechanism [25], inspired by the convolutional block attention module. The attention mechanism is illustrated by the orange box in Fig. 4. The channel attention module recalibrates the significance of features across channels by

Optics EXPRESS

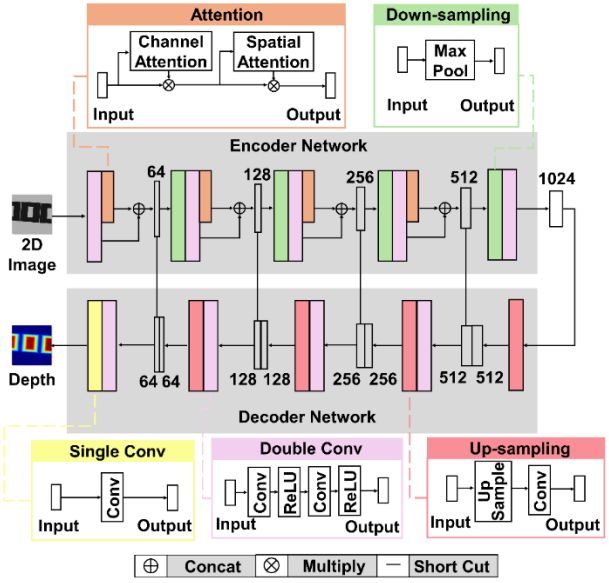

Fig. 4. The architecture of UniDepth model.

synthesizing global spatial information. And the refined feature map is computed, emphasizing channels critical to depth cues. Implemented sequentially following channel attention, the spatial attention module focuses on spatially salient region. It integrates features derived from average-polling and max-polling operations across channels. These attention mechanisms aim to enhance performance by focusing on important features and suppressing less relevant ones. The channel attention module and spatial attention module are applied sequentially to learn which features to emphasize across different dimensions. This approach facilitates data flow within the network by determining which information to highlight or suppress. The down-sampling process employs max-polling, halving the spatial dimensions after each double convolution block while retaining the prominent features. This operation is repeated across five stages, reducing the input resolution and yielding bottleneck feature map. This multi-scale encoding captures both fine-grained details and global context, essential for accurate depth estimation from 2D images.

After the encoder processes the 2D image, its features are extracted for further analysis. The encoder replicates the structure of encoder, progressively up-sampling the bottleneck features to reconstruct a dense depth map. The decoder consists of double convolution layers and up-sampling. The up-sampling operations double the spatial dimensions at each stage, reducing the channel count. And like the encoder, each up-sampling step is followed by a double convolution block refining the fused features, maintaining consistency in feature processing. The decoder also incorporates shortcut connections, features from corresponding encoder stage are concatenated with up-sampled features, preserving low-level details lost during down-sampling. The final layer employs a single convolution to map the output channel to a single channel depth, followed by a sigmoid activation function. This architecture ensures that the spatial resolution of the input is fully restored in the output depth map.

Compared to existing depth estimation methods [26–29], UniDepth delivers more precise depth estimation, resulting in highly accurate and detailed 3D structures. By leveraging muti-scale feature extraction and attention mechanism, UniDepth offers enhanced robustness across diverse scenarios, ultimately producing finer-grained and more faithful 3D reconstructions in measurement. UniDepth model produces a dense depth map, however its prediction results exhibit inherent noises and minor inconsistences. These issues are particularly evident in regions

Optics EXPRESS

characterized by low-texture or complex micro-structure. Consequently, a refinement step is essential to achieve optimal quality. To address these issues, a guided filter [30] is utilized as a post-processing technique to enhance the raw depth output of UniDepth, thereby improving its accuracy for micro-structured surface profiling in the MAM system. The guided filter, an edge-preserving smoothing method, employs the HR image from the 2D optical channel of the MAM system as a guidance image, aligning the depth map with the structural features evident in the data.

The loss function for UniDepth model in this paper comprises several parts: a smooth L1 loss, a perceptual loss, and an image gradient loss. The smooth L1 loss is used to calculate the pixel error between the predicted image and ground truth. This loss combines the advantages of L1 loss and L2 loss, enabling the training process more stable. The smooth L1 loss is formulated as:

$$L_s = \begin{cases} 0.5 \cdot (I_{pre} - I_{gt})^2/\beta \text{ if } |I_{pre} - I_{gt}| < \beta \\ |I_{pre} - I_{gt}| - 0.5 \cdot \beta \text{ if} |I_{pre} - I_{gt}| \geq \beta \end{cases}. \tag{5}$$

where I_{pre} is the predicted depth, I_{gt} is the ground truth corresponding to the predicted depth, β is a parameter that controls the extent of the smoothing area, in this paper β equals to 1. Since the smooth L1 loss solely evaluates the pixel-wise difference, some attributes such as image style and image similarity are disregarded. Consequently, Perceptual loss is employed to avoid low-quality depth and a distortion. Perceptual loss prioritizes the perceived quality of the image, aligning more closely with the human visual perception [31]. To determine perceptual loss, the first 35 layers (convolution layers, activation functions, and pooling) of the VGG19 model [32] are loaded. Although loading these layers of VGG19 indeed increases the training time, it is essential to compute the perceptual loss, as the convolution layers of the network effectively extract structural information. This deep learning architecture facilitates the comparison of feature representations between predicted depth and ground truth. The discrepancy between these representations is quantified by mean squared error (MSE). The perceptual loss is formulated as:

$$L_p = \frac{1}{N} \sum (M(I_{pre}) - M(I_{gt}))^2. \tag{6}$$

where $M(\cdot)$ denotes the process of feature extraction executed by the VGG19 model, with its parameters frozen throughout the training process. N represents the batch size utilized within a single iteration throughout the network training process. The image gradient loss facilitates the preservation of enhanced edge details and high-frequency information within the generated image by quantifying the gradient disparity between the predicted image and the ground truth. The image gradient loss prioritizes the structural attributes of the image over the straightforward application of pixel-level loss. The image gradient loss is formulated as:

$$L_g = \frac{1}{H \cdot W} \sum_{x,y} ((G_x(I_{pre})(x,y) - G_x(I_{gt})(x,y))^2 + (G_y(I_{pre})(x,y) - G_y(I_{gt})(x,y))^2) \tag{7}$$

where H and W are the height and width of the depth, $G_x(\cdot)$ represents the horizontal gradient of the depth, $G_y(\cdot)$ represents the vertical gradient. Sobel is applied for image gradient calculations in this paper.

The final loss function is represented as:

$$L_f = L_s + L_p + L_g. \tag{8}$$

3.1.3. Depth fusion

For the depth fusion process, a strategy based on the theory of pyramid representation is proposed, as illustrated in Fig. 5. Its purpose is to effectively integrate depth information calculated from

Optics EXPRESS

multiple sources in the MAM system. This approach leverages the complementary strengths of two depth maps to mitigate inconsistences such as noises, edge discontinuities, and resolution disparities. Key part to this fusion strategy is the use of an all-in-focus image. This image is generated via autostereoscopic refocusing process. It serves as a reference for comparing the similarities of HF information and LF information between two depth maps. This reference image, rich in spatial detail across all depths, provides a robust baseline for ensuring structural consistency in the fused output.

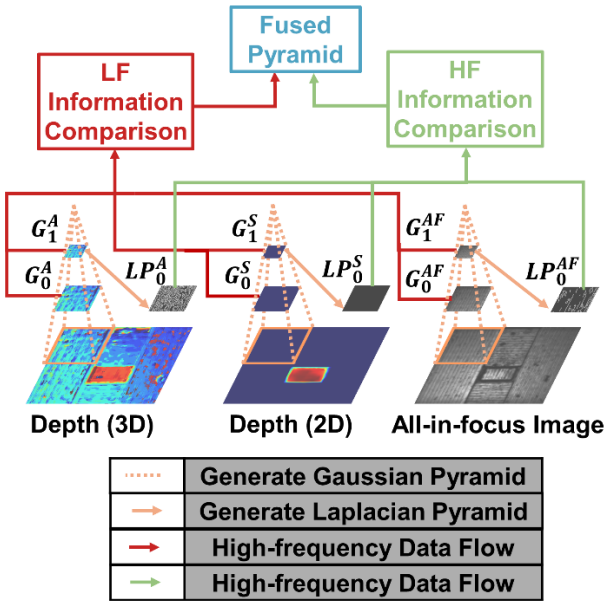

Fig. 5. The procedure of depth fusion.

The fusion process begins with an alignment step, where an affine transformation is performed to register the HR image and the all-in-focus image. Feature points, extracted using a scale-invariant feature transform (SIFT), are matched between the two images to estimate the transformation parameters. This alignment ensures the depth maps are spatially consistent with the reference image, facilitating accurate comparisons across images. To analyze the detailed information of each image and filter out noises while preserving critical structural features, a sliding window mechanism is employed. This window slides systematically along both the horizontal and vertical directions of the depth maps and all-in-focus reference image. During each slide, the ROI is extracted from the aligned images, encompassing corresponding patches from two depth maps and the reference. The ROIs serve as the foundational units for multi-scale analysis and fusion. The extracted ROIs are then decomposed using a Gaussian pyramid representation, which facilitates the separation of HF and LF information at different scales. The Gaussian pyramid is constructed by iteratively applying a Gaussian blur followed by down-sampling to generate a hierarchy of images. The Gaussian pyramid is mathematically represented as:

$$G_l(i,j) = \sum_{m=-2}^{2}\sum_{n=-2}^{2} g(m,n)G_{l-1}(2i+m, 2j+n). \quad (1 \le l < N) \tag{9}$$

where l represents the level of the pyramid, in this paper l equals to 3, $g(m,n)$ is a Gaussian filter, i and j are the number of the column.G_0, G_1 and G_2 constitute the Gaussian pyramid

Optics EXPRESS

representation. In Fig. 5, G_l^A represents Gaussian pyramid level of 3D depth information, G_l^S represents Gaussian pyramid level of 2D depth information, G_l^{AF} represents Gaussian pyramid level of all-in-focus image. Based on the Eq. (9), the present layer of the Gaussian pyramid is formed through the application of Gaussian filtering and down-sampling to the previous layer, considering both rows and columns. After the generation of the Gaussian pyramid, Laplacian pyramid is formulated through the following expressions:

$$\begin{cases} LP_l = G_l - G_{l+1}^* & (0 \le l < N) \\ LP_N = G_N & (l = N) \end{cases}. \tag{10}$$

$$G_l^*(i,j) = 4 \sum_{m=-2}^{2} \sum_{n=-2}^{2} g(m,n) G_{l-1}\left(\frac{i+m}{2}, \frac{j+n}{2}\right). \quad (0 < l \le N) \tag{11}$$

where N is the number of the top layer of the Laplacian pyramid, LP_l is the l layer of the Laplacian pyramid, LP_0, LP_1 and LP_1 constitute Laplacian pyramid. In Fig. 5, LP_l^A represents Laplacian pyramid level of 3D depth information, LP_l^S represents Laplacian pyramid level of 2D depth information, LP_l^{AF} represents Laplacian pyramid level of all-in-focus image. After the composition, three groups of image pyramids are obtained. For each ROI, the pyramid decomposes the depth maps and the all-in-focus image into multi-scale representations. The Gaussian pyramid contains LF information at different scales, and the Laplacian pyramid contains HF information. This decomposition enables a detailed comparison of HF and LF information. The fusion strategy then evaluates the similarity between the corresponding pyramid levels of the two depth maps and the reference using the structural similarity index measure (SSIM). Specifically, for each level, the HF and LF information from the depth maps against the pyramid levels of reference, with the one exhibiting greater similarity retained for the fused pyramid. This selective fusion ensures that the final depth map preserves the most reliable and structurally consistent features from multiple sources. The process concludes by reconstructing the fused depth map from the combined pyramid using an inverse pyramid transformation, where the LF base is iteratively up-sampled and combined with the HF details from lower levels. Based on Eq. (10), the inverse pyramid transformation is represented as:

$$\begin{cases} G_N = LP_N & (l = N) \\ G_l = LP_l + G_{l+1}^* & (0 \le l < N) \end{cases}. \tag{12}$$

After the inverse pyramid transformation, the coarse depth map is acquired. In the refinement process, a guided filter is employed. This filter enables preserving the edge information and filtering the noises in images. This multi-scale fusion approach not only enhances the robustness of depth estimation within the MAM system but also leverages the complementary nature of autostereoscopic and 2D optical data, yielding a refined depth map with improved edge definition and reduced noise, critical for applications such as micro-surface reconstruction and 3D visualization.

4. System setup and experiments

4.1. System setup and network learning details

To illustrate the enhancements over previous study, a micro-structured sample was employed to ensure control over experimental variables. The sample consisted of a surface featuring frustum-shaped microstructures. The sample was fixed on a precision three-axis positioning

Optics EXPRESS

stage, enabling controlled lateral and longitudinal movement. The designed measurement system and platform is shown in Fig. 6. The entire system primarily comprises an objective lens, co-axial illumination, an MLA and two image sensors. The objective lens, along with the zoom system, facilitates the magnification of the measured sample. Co-axial illumination adjusts the brightness of the captured images. A beam splitter divides the optical channel into two parts. The proposed system incorporates two CCD sensors. The sensor utilized in the 2D optical channel enables capturing HR images. The sensor employed in the 3D optical channel is designed to detect objects in different perspectives. An MLA is placed in front of the sensor in 3D optical channel. The hardware components for the MAM system are selected to meet the requirements of high-precision imaging and geometric accuracy. The hardware components for the MAM system are selected to meet the requirements of high-precision imaging and geometric accuracy. The selection criteria are based on performance parameters, compatibility with system design, and robustness under experimental conditions. Table 1 shows the specifications of the system.

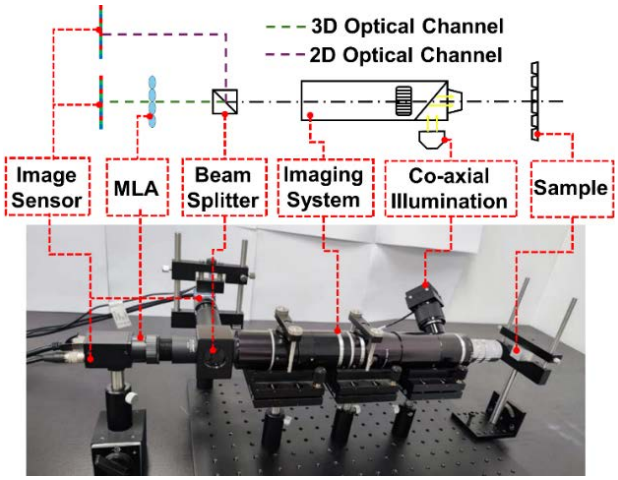

Fig. 6. The setup of the MAM system.

MAM system incorporates two optical channels, and its performance calculations are detailed based on the principles of microscopy and 3D imaging [33]. The magnification of the system is first determined and denoted as:

$$M = \frac{f_t}{f_o} \tag{13}$$

where f_t represents the focal length of tube lens, f_o represents the focal length of objective lens. The horizontal FOV is subsequently calculated. For digital imaging, the FOV can be determined by:

$$FOV = \frac{R_h \Delta p}{M} \tag{14}$$

where R_h is the horizontal resolution of the CCD sensor, Δp is the pixel size of the CCD sensor. The optical resolution of the system is estimated using the Abbe diffraction limit, where the resolution is defined as:

$$R = \frac{0.61\lambda}{NA} \tag{15}$$

where λ is the wavelength of illumination light, NA is the numerical aperture of the objective lens. The depth of field (DOF), representing the distance of an object that remains in focus, is

Optics EXPRESS

Table 1. Specifications of the MAM system

Item	Specification	
CCD Sensor (2D)	Manufacturer	Hikrobot
	Model Name	MV-CA050-10GC
	Pixel Size	3.45 μm
	Sensor Size	2/3 inch
	Resolution	2456 × 2058
CCD Sensor (3D)	Manufacturer	JoinHope
	Model Name	OK_AC5067
	Pixel Size	3.45 μm
	Sensor Size	2/3 inch
	Resolution	2448 × 2048
MLA	Manufacturer	Edmund
	Stock Number	#64-479
	Pitch	500 μm
	Focal Length	13.8 mm
	Scale	10 × 10
Imaging System	Manufacturer	Navitar
	NA	0.28
	Magnification	10×
	Adjustable Zoom	1-12

calculated. The DOF can be approximated as:

$$DOF = \frac{\lambda n}{NA^2} \tag{16}$$

where n is the refractive index of the medium. For the autostereoscopic imaging system in the 3D channel, the FOV is represented as:

$$FOV_{3D} = \frac{p}{M_{3D}} \tag{17}$$

where p is the pitch of MLA, M_{3D} is the magnification ratio of the 3D channel. The NA of the MLA is formulated as:

$$NA_{mla} = \frac{p}{2f_{mla}} \tag{18}$$

where f_{mla} is the focal length of MLA. Since NA_{mla} is significantly smaller than NA_{obj}, the effective NA of the whole system is limited by MLA. In the 2D channel of the MAM system, the FOV is 706.1 μm, with an achievable resolution of 1.2 μm and a DOF of 7.1 μm. In the 3D channel of the MAM system, the FOV is 294.1 μm, with an achievable resolution of 16.8 μm and a DOF of 1399.6 μm.

High precision calibration and alignment of optical systems are critical steps in ensuring optimal performance. This is particularly true for high-precision measurement systems. In this study, the calibration process of autostereoscopic systems is shown in Fig. 7, where a standard target serves as the calibration reference. This target is mounted on an XY translation stage and moved axially within the depth of field of the MAM system. As the target position varies, multiple calibration data points are acquired to facilitate disparity extraction through digital refocusing. By using multiple calibration data, a curve that maps the relationship between disparity and

Optics EXPRESS

moving distance of the reference target can be fitted for the calibration of the system. The purpose of alignment is to ensure that the optical axes and focal planes of each component in the optical system are precisely matched to maximize system performance. The common methods include mechanical alignment, laser alignment, spot method and interferometric alignment. Since the MAM system consists of multiple tube-shape optical components, alignment is mainly focused on the MLA and CCD sections. In this study, laser alignment is employed to align components in the sharing optical path of MAM system.

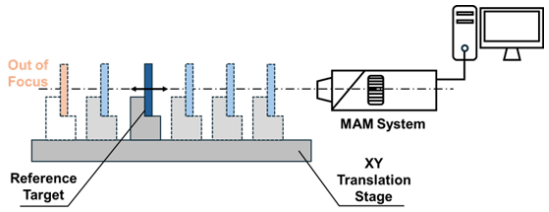

Fig. 7. Calibration process of the MAM system.

The dataset of the real-word micro-structured surfaces currently remains insufficient to support robust training of the UniDepth deep learning model. This insufficiency arises primarily due to the limited availability of high-quality samples and the labor-intensive process of acquiring ground truth depth maps. To address this challenge and ensure effective model training, a simulated dataset tailored to represent common micro-structured surfaces is generated using Blender. This synthetic dataset encompasses a variety of representative micro-structures frequently encountered in industrial and scientific applications, including frustums, spherical holes, and V-grooves. Each category comprises 625 images, with each image rendered at a resolution of 512×512 pixels. To enhance the robustness and generalization of the UniDepth model, data augmentation techniques are applied prior to training. These include translation, rotation, and flip were executed. These augmentations triple the effective dataset size, enriching the exposure of the model to diverse spatial configurations. Furthermore, diverse lighting conditions and measured sample positions are simulated in Blender. The variety of measured samples is increased to accommodate different measurement tasks. During the generation of synthetic data, random errors are intentionally introduced to simulate the inaccuracy of the real-world.

In the depth fusion process, the selection of the sliding window affects the depth fusion results. A large sliding window effectively captures local features, such as edges and textures, but excessively large windows may lead to loss of detail or blurred boundaries. In this paper, the size of the sliding window is estimated by resolution and physical dimensions of the measured sample. The approximate range of window size is between 8 pixels to 30 pixels. The fusion results are shown in Fig. 8. We tested different window sizes, evaluating their impact on fused depth map accuracy. After multiple experiments, the results demonstrate that an 8-pixel sliding window yields optimal performance.

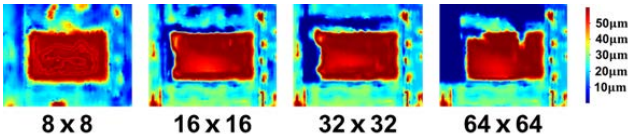

Fig. 8. The fusion results based on different sliding windows.

To further prevent overfitting and improve resilience to real-word imperfections, noise injections are incorporated into the simulated dataset. Gaussian noise is added to simulated 2D images,

Optics EXPRESS

mimicking sensor noise and minor surface irregularities typical of physical captures. Additionally, random speckle noise is introduced to simulate optical artifacts ensuring the model learns to distinguish structural features from noise-induced distortions. This noise-augmented dataset better approximates the challenges encountered in the real-world deployment of MAM system, such as lighting variations and sensor limitations. The training pipeline integrates this synthetic dataset with a split of 80% for training, 10% for validation, and 10% for testing, stratified across the three micro-structure categories to maintain balance. UniDepth model is implemented in Pytorch. The ReLU were employed as the activation functions for the UniDepth model. The Adam optimizer was employed in the training process. The total number of training epoch is 500. The initial value of deep learning rate was 10^{-4}, with the learning rate of each parameter group decayed by a factor of 0.1 every 100 epochs. This model was trained on a Nvidia GeForce RTX 2080 graphics card and equipped with an Intel Core i7-8700 central processing unit.

For depth fusion, a three-layer Gaussian pyramid and two-layer Laplacian pyramid were constructed for each sliding window operation. The size of the sliding window is specified as 8×8. The window moves with a stride of 1.

4.2. Experiments and analysis

All the experiment results were obtained in accordance with the system as shown in Fig. 6. The acquired EIs went through digital refocusing, and reconstruction. The digital refocusing process was based on the method proposed in [10] and then reconstruction was conducted using the approach proposed in [34]. The UniDepth model generated dense depth maps via acquired 2D images. Through the pyramid representation the geometry of the measured surfaces was obtained. The input data and depth estimation results are shown in Fig. 9. To comprehensively evaluate the effectiveness and novelties of the proposed method within the MAM system, this experimental study presents qualitative and quantitative comparisons.

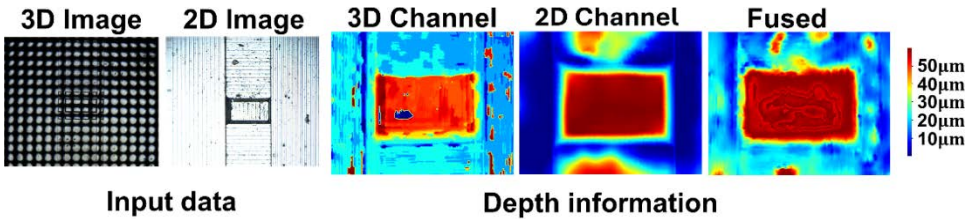

Fig. 9. The input data and depth estimation results.

To explore the advancement of the UniDepth, qualitative and quantitative experiments are conducted with existing deep learning methods [26–29]. StructDepth [26] is a self-supervised monocular depth estimation method that utilizes two additional supervisory signals for training. MiDaS [27] is a robust monocular depth estimation model trained on a composite of multiple datasets. DPT [28] employs vision transformers as a backbone for dense prediction tasks. LeReS [29] adopts a two-stage framework that initially predicts depth from a single monocular image up to an unknown scale and shift. These methods collectively represent significant advancements in deep learning-based depth estimation. Figure 10 shows the qualitative analysis between different methods under the test dataset. Table 2 shows the quantitative experiment under peak signal-to-noise ratio (PSNR) and SSIM. These experiments show that UniDepth achieves more accurate depth estimation and produces more accurate 3D structures, compared with the existing methods.

To explore the error propagation, error analysis in different channels is conducted. Accuracy was quantified by calculating the root mean square error (RMSE) and maximum absolute error (MaxAE) between the aligned measurement data and reference values for each channel across

Optics EXPRESS

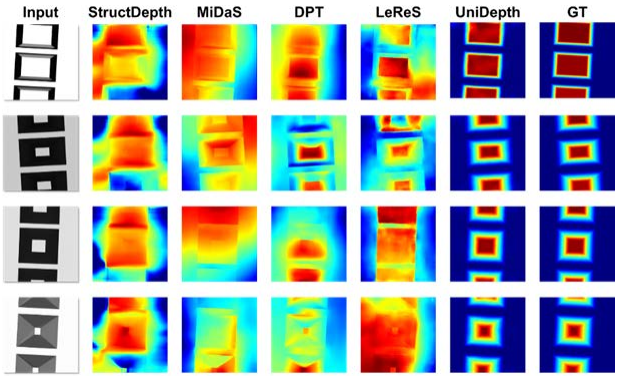

Fig. 10. Qualitative analysis between different deep learning methods.

Table 2. Quantitative analysis between different deep learning methods

	StructDepth	MiDaS	DPT	LeReS	UniDepth
PSNR ↑	6.24	5.90	5.68	6.70	**29.29**
SSIM ↑	0.2832	0.2760	0.2790	0.2884	**0.9611**

multiple trials. As shown in Fig. 11, the fusion process leverages the complementarity of different depth information, integrating the strengths of each channel to mitigate the limitations of individual channels.

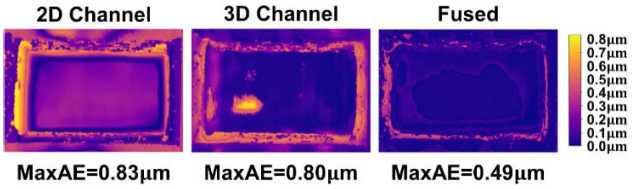

Fig. 11. Error analysis in different channels.

A qualitative comparison study is presented to evaluate the fusion results of the proposed MAM system against traditional autostereoscopic measurement systems employing a variety of established depth estimation methods. In this paper, the shape from focus algorithm based on Laplacian operator, EPI-shift [35], FastLFnet [36] based on cost volume and super resolution method SSA [34] are applied in the traditional autostereoscopic system. These methods are selected to represent a spectrum of autostereoscopic 3D reconstruction strategies, ranging from focus-based, disparity-based and deep learning -based approaches. The visual comparison is shown in Fig. 12, with outcomes from the Zygo Nexview Optical Profiler serving as a reference.

In the traditional autostereoscopic system, each method exhibits distinct limitations, particularly when applied to the real-word sample. The EPI-shift method struggles to find unique displacement matches in low-texture regions, resulting in noises or errors. Local matching in EPI method blurs object boundaries causing edge discontinuities marked by the black box. The shape from focus method is highly sensitive to high-frequency image information like edges and textures, causing several artifacts, as marked by the black box in the depth map. While the FastLFnet method has been adopted for cost aggregation in non-edge regions, this approach fails to preserve the accuracy and generate excessive noises in non-edge regions. The SSA method employs generative network to generate novel perspectives. It enhances 3D construction compared to the

Optics EXPRESS

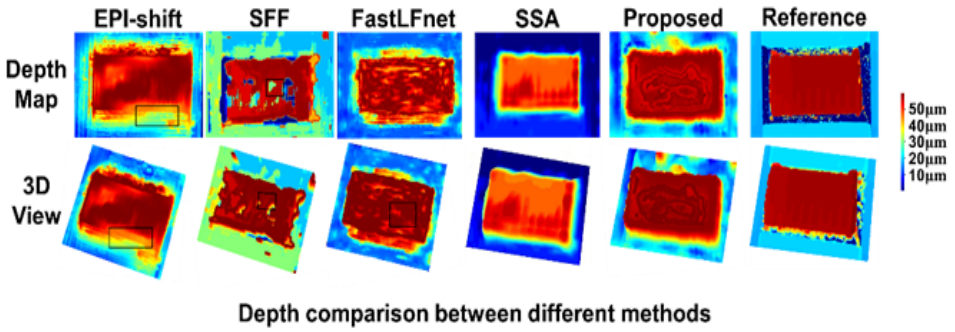

Fig. 12. The qualitive comparison between EPI-shift, SFF, FastLFnet, SSA, proposed method and reference.

traditional autostereoscopic methods. In contrast, the depth map and 3D view from the proposed method exhibit a closer resemblance to the reference, particularly in terms of reconstructing the edges and surfaces. The proposed method adopts a fusion strategy to compensate for the limitations of previous methods in low-texture regions, balances noises suppression and details preservation. During fusion progress, this system retains optimized HF information and LF information, particularly in precisely reconstructing edges and surfaces.

To assess the repeatability and uncertainty analysis of the proposed MAM system against traditional autostereoscopic system, a total of 10 repeated measurements were performed on a representative micro-structured sample. This experiment compares EPI-shift, shape from focus based on Laplacian operator, FastLFnet based on cost volume, SSA method and the proposed method. The sample was measured under controlled conditions to minimize external variables, with each method applied to the same set of EIs and HR 2D images captured by the dual channels of the MAM system. To align the 3D measurement results with a high-precision reference and enable evaluation, the iterative closest point (ICP) was employed. This algorithm registered the point clouds generated by each method against the ground truth data obtained from the Zygo. Accuracy was quantified by calculating RMSE and MaxAE are determined between the aligned measurement data and the reference for each method across the trials. As shown in Fig. 13(a), it is observed that the 3D measurement results obtained from the proposed system exhibit strong agreement with those measured using Zygo system.

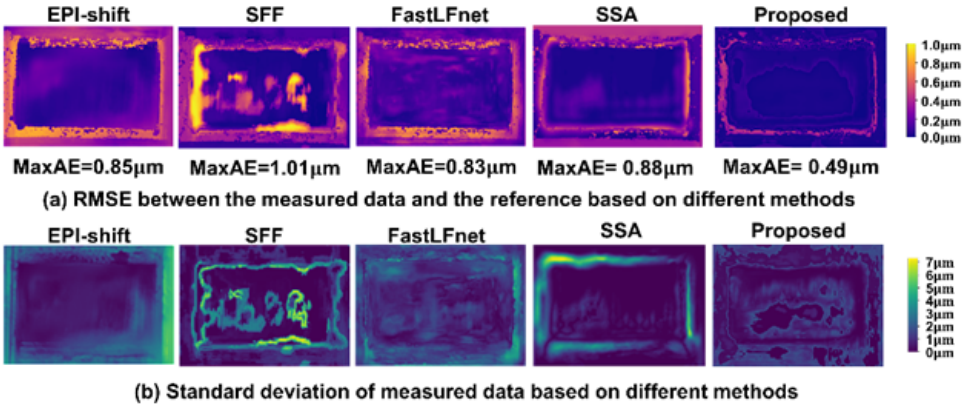

Fig. 13. The quantitative comparison between EPI-shift, SFF, FastLFnet, SSA, proposed method and reference, (a) RMSE and MaxAE, (b) standard deviation.

Optics EXPRESS

The standard deviation of the measurement results has been calculated, and the values are represented in Fig. 13(b). The shape from focus method based on Laplacian operator is unable to produce smooth surfaces. The FastLFnet and EPI-shift exhibits fewer surface flaws, while a substantial standard deviation is obtained. Although, the SSA method demonstrates visually appealing outcomes with well-defined geometric shapes and consistent surface textures, the quantitative evaluation reveals that the reconstruction error remains relatively high, primarily due to noise introduced from the inaccurate new-view synthesis. The SSA method does not consistently produce accurate disparities, resulting in blurriness in refocused images. The proposed system has small standard deviation and smooth surface, which means the multichannel system has the capacity of consistent and robust measurements.

5. Conclusion

To address the limitations inherent in traditional autostereoscopic systems, this study set out to establish a multichannel measurement system to measure the micro-structured surfaces. The system comprises a 3D optical channel designed to capture EIs from multiple viewpoints, paired with a 2D channel dedicated to acquiring HR images. For the MAM approach, a deep learning model was utilized to estimate the 3D geometry from HR images, while autostereoscopic technology was employed to obtain depth information from EIs. By integrating depth information from multiple channels using a pyramid representation, the system leverages their complementary attributes while effectively mitigating noises and derives an optimal depth. The experimental results demonstrate that this novel measurement system achieves consistent and robust measurement in depth estimation, outperforming traditional autostereoscopic systems employing different methods. For further research, the MAM system will be enhanced by integrating multiple optical routes including micro zoom, wide-angle and comprehensive 3D measurement data.

Funding. Research Grants Council of the Government of the Hong Kong Special Administrative Region (R5047-22, 15207521); Research Committee of the Hong Kong Polytechnic University.

Disclosures. The authors declare no conflicts of interest.

Data availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

References

1. Z. Fang, H. Zhang, X. Zhou, *et al.*, "Design and fabrication of a controllable haze diffuser film," in *LED and Display Technologies II*, 2012, 8560: SPIE, pp. 32–37.
2. L. Jia, J. Jiang, T. Xiang, *et al.*, "Multifunctional biomimetic microstructured surfaces for healthcare applications," Adv. Mater. Interfaces **9**(33), 2201270 (2022).
3. J.-F. Seurin, D. Zhou, G. Xu, *et al.*, "High-efficiency VCSEL arrays for illumination and sensing in consumer applications," in *Vertical-Cavity Surface-Emitting Lasers XX*, 2016, 9766: SPIE, pp. 60–68.
4. S. Carmignato, L. De Chiffre, H. Bosse, *et al.*, "Dimensional artefacts to achieve metrological traceability in advanced manufacturing," CIRP Ann. **69**(2), 693–716 (2020).
5. S. Yang and G. Zhang, "A review of interferometry for geometric measurement," Meas. Sci. Technol. **29**(10), 102001 (2018).
6. J. Fan and F. Wang, "Review of ultrasonic measurement methods for two-phase flow," Rev. Sci. Instrum. **92**(9), 091502 (2021).
7. W. Yin, S. Feng, T. Tao, *et al.*, "High-speed 3D shape measurement using the optimized composite fringe patterns and stereo-assisted structured light system," Opt. Express **27**(3), 2411–2431 (2019).
8. S. Li, X. Jia, M. Chen, *et al.*, "Error analysis and correction for color in laser triangulation measurement," Optik **168**, 165–173 (2018).
9. D. Li, C. F. Cheung, M. Ren, *et al.*, "Autostereoscopy-based three-dimensional on-machine measuring system for micro-structured surfaces," Opt. Express **22**(21), 25635–25650 (2014).
10. D. Li, C. F. Cheung, M. Ren, *et al.*, "Disparity pattern-based autostereoscopic 3D metrology system for in situ measurement of microstructured surfaces," Opt. Lett. **40**(22), 5271–5274 (2015).
11. J. Li, M. Lu, and Z.-N. Li, "Continuous depth map reconstruction from light fields," IEEE Transactions on Image Processing **24**(11), 3257–3265 (2015).
12. W. Williem and I. K. Park, "Robust light field depth estimation for noisy scene with occlusion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4396–4404.

Optics EXPRESS

13. H.-G. Jeon, J. Park, G. Choe, *et al.*, "Accurate depth map estimation from a lenslet light field camera," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1547–1555.
14. S. Heber and T. Pock, "Convolutional networks for shape from light field," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3746–3754.
15. S. Gao and C. F. Cheung, "Autostereoscopic 3D measurement based on adaptive focus volume aggregation," Sensors **23**(23), 9419 (2023).
16. X. Qin, Z. Zhang, C. Huang, *et al.*, "U2-Net: Going deeper with nested U-structure for salient object detection," Pattern Recognition **106**, 107404 (2020).
17. T. Wang, A. Borji, L. Zhang, *et al.*, "A stagewise refinement model for detecting salient objects in images," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4019–4028.
18. J.-J. Liu, Q. Hou, M.-M. Cheng, *et al.*, "A simple pooling-based design for real-time salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3917–3926.
19. H. Zhao, J. Shi, X. Qi, *et al.*, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.
20. K. Ashfaq and M. T. Mahmood, "Enhancing focus volume through perceptual focus factor in shape-from-focus," Mathematics **12**(1), 102 (2023).
21. U. Ali, I. H. Lee, and M. T. Mahmood, "Guided image filtering in shape-from-focus: A comparative analysis," Pattern Recognition **111**, 107670 (2021).
22. C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 270–279.
23. D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in neural information processing systems*, 27, 2014.
24. D. Xu, W. Wang, H. Tang, *et al.*, "Structured attention guided convolutional neural fields for monocular depth estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3917–3925.
25. S. Woo, J. Park, J.-Y. Lee, *et al.*, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.
26. B. Li, Y. Huang, Z. Liu, *et al.*, "StructDepth: Leveraging the structural regularities for self-supervised indoor depth estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12663–12673.
27. R. Ranftl, K. Lasinger, D. Hafner, *et al.*, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," IEEE Trans. Pattern Anal. Mach. Intell. **44**(3), 1623–1637 (2022).
28. R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12179–12188.
29. W. Yin, J. Zhang, O. Wang, *et al.*, "Learning to recover 3d scene shape from a single image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 204–213.
30. K. He, J. Sun, and X. Tang, "Guided image filtering," IEEE Trans. Pattern Anal. Mach. Intell. **35**(6), 1397–1409 (2013).
31. J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, 2016: Springer, pp. 694–711.
32. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv 2014.
33. M. Martínez-Corral and B. Javidi, "Fundamentals of 3D imaging and displays: a tutorial on integral imaging, light-field, and plenoptic systems," Adv. Opt. Photonics **10**(3), 512–566 (2018).
34. S. Gao, C. F. Cheung, and D. Li, "Self super-resolution autostereoscopic 3D measuring system using deep convolutional neural networks," Opt. Express **30**(10), 16313–16329 (2022).
35. T. Leistner, H. Schilling, R. Mackowiak, *et al.*, "Learning to think outside the box: Wide-baseline light field depth estimation with EPI-shift," in *2019 international conference on 3D vision (3DV)*, 2019: IEEE, pp. 249–257.
36. Z. Huang, X. Hu, Z. Xue, *et al.*, "Fast light-field disparity estimation with multi-disparity-scale cost aggregation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6320–6329.