Opto-Electronic Science

CN 51-1800/O4 ISSN 2097-0382 (Print) ISSN 2097-4000 (Online)

Edge enhanced depth perception with binocular meta-lens

Xiaoyuan Liu, Jingcheng Zhang, Borui Leng, Yin Zhou, Jialuo Cheng, Takeshi Yamaguchi, Takuo Tanaka and Mu Ku Chen

Citation: Liu XY, Zhang JC, Leng BR, et al. Edge enhanced depth perception with binocular meta-lens. *Opto-Electron Sci* **3**, 230033 (2024).

https://doi.org/10.29026/oes.2024.230033

Received: 26 September 2023; Accepted: 18 December 2023; Published online: 2 April 2024

Related articles

Improved spatiotemporal resolution of anti-scattering super-resolution label-free microscopy via synthetic wave 3D metalens imaging

Yuting Xiao, Lianwei Chen, Mingbo Pu, Mingfeng Xu, Qi Zhang, Yinghui Guo, Tianqu Chen, Xiangang Luo

Opto-Electronic Science 2023 **2**, 230037 doi: 10.29026/oes.2023.230037

Multi-foci metalens for spectra and polarization ellipticity recognition and reconstruction

Hui Gao, Xuhao Fan, Yuxi Wang, Yuncheng Liu, Xinger Wang, Ke Xu, Leimin Deng, Cheng Zeng, Tingan Li, Jinsong Xia, Wei Xiong

Opto-Electronic Science 2023 **2**, 220026 doi: 10.29026/oes.2023.220026

Dynamic phase assembled terahertz metalens for reversible conversion between linear polarization and arbitrary circular polarization

Jitao Li, Guocui Wang, Zhen Yue, Jingyu Liu, Jie Li, Chenglong Zheng, Yating Zhang, Yan Zhang, Jianquan Yao

Opto-Electronic Advances 2022 **5**, 210062 doi: 10.29026/oea.2022.210062

Ultrahigh performance passive radiative cooling by hybrid polar dielectric metasurface thermal emitters

Yinan Zhang, Yinggang Chen, Tong Wang, Qian Zhu, Min Gu

Opto-Electronic Advances 2024 , doi: 10.29026/oea.2024.230194

More related article in Opto-Electronic Journals Group website

Opto-Electronic Science

http://www.oejournal.org/oes

OE_Journal

Website

Opto-Electronic Science

Special Issue on Meta-lenses

Article

September 2024, Vol. 3, No. 9

DOI: 10.29026/oes.2024.230033

Edge enhanced depth perception with binocular meta-lens

Xiaoyuan Liu[1,2,3], Jingcheng Zhang[1], Borui Leng[1], Yin Zhou[1], Jialuo Cheng[1], Takeshi Yamaguchi[4,5,6], Takuo Tanaka[4,5,6]* and Mu Ku Chen[1,2,3]*

The increasing popularity of the metaverse has led to a growing interest and market size in spatial computing from both academia and industry. Developing portable and accurate imaging and depth sensing systems is crucial for advancing next-generation virtual reality devices. This work demonstrates an intelligent, lightweight, and compact edge-enhanced depth perception system that utilizes a binocular meta-lens for spatial computing. The miniaturized system comprises a binocular meta-lens, a 532 nm filter, and a CMOS sensor. For disparity computation, we propose a stereo-matching neural network with a novel H-Module. The H-Module incorporates an attention mechanism into the Siamese network. The symmetric architecture, with cross-pixel interaction and cross-view interaction, enables a more comprehensive analysis of contextual information in stereo images. Based on spatial intensity discontinuity, the edge enhancement eliminates ill-posed regions in the image where ambiguous depth predictions may occur due to a lack of texture. With the assistance of deep learning, our edge-enhanced system provides prompt responses in less than 0.15 seconds. This edge-enhanced depth perception meta-lens imaging system will significantly contribute to accurate 3D scene modeling, machine vision, autonomous driving, and robotics development.

Keywords: metasurfaces; meta-lenses; deep learning; depth perception; edge detection

Liu XY, Zhang JC, Leng BR et al. Edge enhanced depth perception with binocular meta-lens. *Opto-Electron Sci* **3**, 230033 (2024).

Introduction

Spatial computing[1] and the emerging meta-verse represent a paradigm shift in how humans interact with a machine. Spatial computing refers to integrating digital information and virtual objects into the physical world, creating a mixed reality where the boundaries between the digital and physical realms are blurred. Common augmented reality devices rely on spatial computing to perceive the depth of the real physical world while embedding virtual objects into real scenes three-dimensionally[2]. One of the key technologies of spatial computing is its depth perception capability, which bridges the gap between the physical and digital realms. This promises intuitive and natural interaction with virtual objects. Therefore, digital information can be correctly placed and manipulated in the scene following physical laws. However, the weight and volume of traditional depth

[1]Department of Electrical Engineering, City University of Hong Kong, Hong Kong SAR 999077, China; [2]Centre for Biosystems, Neuroscience, and Nanotechnology, City University of Hong Kong, Hong Kong SAR 999077, China; [3]The State Key Laboratory of Terahertz and Millimeter Waves, and Nanotechnology, City University of Hong Kong, Hong Kong SAR 999077, China; [4]Innovative Photon Manipulation Research Team, RIKEN Center for Advanced Photonics, 351-0198, Japan; [5]Metamaterial Laboratory, RIKEN Cluster for Pioneering Research, 351-0198, Japan; [6]Institute of Post-LED Photonics, Tokushima University, 770-8506, Japan.

*Correspondence: T Tanaka, E-mail: t-tanaka@riken.jp; MK Chen, E-mail: mkchen@cityu.edu.hk

Received: 26 September 2023; Accepted: 18 December 2023; Published online: 2 April 2024

sensing systems result in a lack of comfort in human-computer interaction wearable devices, which contain many sensors (mainly cameras and LiDAR). At the same time, the space occupied by bulky sensors also limits battery life, causing the device to need to be recharged frequently. Advancements in portable and accurate imaging and depth sensing systems are crucial for next-generation human-computer interaction wearable devices.

Complementing spatial computing, binocular meta-lens[3] offers a breakthrough approach to depth sensing and imaging with the advantages of being lightweight[4, 5], thin, and compact. Meta-lens create advanced optical functionalities that surpass the limitations of traditional optics[6, 7], such as wavefront shaping[8], polarization control[9-11], and spectral manipulation[12, 13]. Meta-lens utilizes nanoantennas to manipulate light[14], offering an opportunity for engineering optical properties such as thinness, flatness, broadband capability[15], high diffraction efficiency[16], extreme depth-of-field[17], and compatibility with complementary metal-oxide-semiconductor (CMOS) technology. By leveraging the unique properties of meta-optics, this compact and miniaturized optical meta-device allows for capturing three-dimensional information from the surrounding environment. Binocular meta-lens enable precise and accurate depth perception, similar to human binocular vision. In recent years, the support of artificial intelligence has increasingly promoted the development of meta-devices in terms of inverse design[18, 19], prompt data analysis[20, 21], optical computation[22, 23], and intelligent reconfigurable meta-devices[24, 25]. These advancements pave the way for compact, lightweight, and highly efficient optical systems seamlessly integrated into spatial computing devices, enhancing their performance and enabling novel applications.

The principle underlying depth acquisition in binocular imaging relies on presenting a stereo-image pair exhibiting discernible disparities[26]. Disparity denotes the horizontal displacement between corresponding pixels in the left and right images. Traditional binocular disparity computation pipeline often entails the utilization of block matching algorithms for calculating matching losses[27]. The combination of deep learning and photonics has been widely researched in recent years, encompassing applications such as orbital angular momentum communication[28], optical neural networks[29], optical encryption[30], enhancing holographic data storage (HDS)[31], photonic inverse design[32] and hyperspectral imaging[33]. Nonetheless, convolutional neural networks (CNNs) have garnered greater preference owing to their inherent advantages of rapidity, precision, and operational simplicity in processing. Despite significant advancements in accuracy and speed achieved by various binocular stereo systems, finding accurate corresponding points within inherently ill-posed regions for depth computation remains challenging, such as textureless areas and reflective surfaces[34]. Ambiguous depth prediction has a serious impact on subsequent machine decision-making. Edge is the typical representation of texture. There must be texture feature points in the edge area for stereo matching. Numerous studies have explored edge detection techniques utilizing meta-lenses, each with distinct characteristics. For instance, the Green function[35, 36], and spiral phase[37] have been employed to enable edge detection using a single meta-lens. Another approach involves utilizing meta-lens arrays for three-dimensional (3D) edge detection[38]. Polarization control has been leveraged for switchable bright field imaging and edge detection capabilities[39, 40]. Edge detection by the Pancharatnam–Berry phase[41] has emerged as a noteworthy technique, demonstrating potential in quantum applications[42]. Edge-based depth perception offers superior fidelity in the estimation of depth. Within the framework of depth edge views, non-textured regions that lack prominent edges or transitions are efficiently discarded. This filtering process reduces the impact of unreliable or ambiguous depth information originating from textureless regions, thereby enhancing the overall accuracy and reliability of depth estimation. By focusing on edges that signify depth discontinuities, edge-based depth perception provides a more resilient and accurate depiction of depth.

We develop an edge-enhanced depth perception based on binocular meta-lens for spatial computing. The whole system is miniaturized, intelligent, lightweight and compact. Its physical working mechanism consists of a binocular lens, a 532 nm filter, and a CMOS sensor. Each meta-lens, measuring 2.6 mm in diameter, weighs 2.45×10^{-5} g and occupies a volume of 3.98×10^{-6} cm^3. The weight of the Sapphire substrate is 0.115 g with a volume of 0.0288 cm^3. Thin and flat nature make it simple in both physical system configuration and image processing pipeline. Without preprocessing, the raw captured image is processed directly by our proposed pyramid stereo-matching neural network, H-Net, to obtain the disparity. A novel symmetric H-module with an attention mechanism allows the H-Net to dynamically

allocate resources based on the significance of contextual features of each view and the correlation between the left and right views. With depth-sensing results, an edge enhancement is performed to filter the feature information that detects the 3D space gradients.

Figure 1 demonstrates the edge-enhanced depth perception system schematic with our binocular meta-lens. There are two letter objects in front of the binocular stereo-vision meta-lens. The application scenario shown in Fig. 1 has ill-posed regions, such as the letter objects' unpatterned backgrounds and untextured surfaces. But with the support of a proposed neural network for comprehensive context analysis and a Canny edge detector for filtering, an edge-enhanced depth perception view is realized, perceiving both intensity and depth discontinuities simultaneously.

The convergence of spatial computing and meta-optics holds immense potential for transforming our daily lives. From augmented and virtual reality experiences that blend seamlessly with our physical surroundings to smart glasses that provide personalized information overlays, edge-enhanced spatial computing powered by meta-optics promises to revolutionize how we perceive and interact with the world around us. This integration can lead to breakthroughs in robotics, autonomous systems, underwater exploration, and medical imaging, where accurate depth perception is crucial for navigation, object recognition, and scene reconstruction.

Methods

Simulation and fabrication

We utilize the commercial simulation software COMSOL Multiphysics® to design and analyze the unit cells of the meta-lens. We set periodic boundary conditions for the x and y directions and a perfect match layer (PML) boundary condition for the z-direction. The meta-lens consists of unit cells of gallium nitride (GaN) cylindrical nanopillars on a sapphire substrate. The diameter of the nanopillars varies across the meta-lens. The refractive index of the sapphire substrate is set to 1.77, while the refractive index of GaN at the working wavelength is 2.42. Using this configuration, we calculate the cylindrical nanopillars' simulated transmission spectra and phase shift, as shown in Supplementary information Fig. S1. The meta-atom arrangement layout for fabrication is designed according to the focusing phase distribution

$$\varphi(x,y,\lambda) = -\left[\frac{2\pi}{\lambda}\left(\sqrt{x^2+y^2+f^2}-f\right)\right], \quad (1)$$

in which $\varphi(x,y,\lambda)$ is the phase compensation requirement at the (x,y) position under the illumination of wavelength $\lambda = 532$ nm, f is the desired focal length of 10.0 mm. The target diameter of each meta-lens is 2.6 mm.

The proposed binocular meta-lens is fabricated by

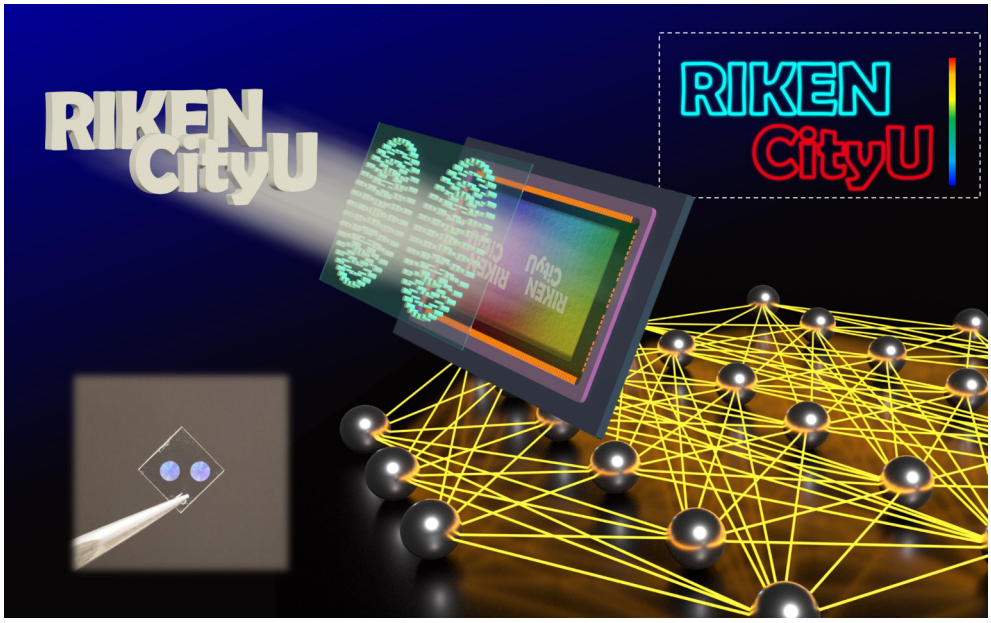

Fig. 1 | Schematic of the edge-enhanced spatial computing with binocular meta-lens. There are two letter objects in front of the binocular meta-lens, which are texture-less and have no background. A binocular meta-lens is designed and fabricated to develop the stereo vision system for texture-less spatial computing scenarios. An edge-enhanced depth perception is realized with the support of a proposed neural network.

adopting the following process (see details in Supporting Information Fig. S2): A 750-nm-thick GaN is firstly deposited on a sapphire substrate via metalorganic chemical vapor deposition (MOCVD). A 200-nm-thick SiO_2 film, which serves as the hard mask for pattern transfer to the GaN layer with a high aspect ratio, is subsequently deposited using an E-gun evaporator. A PMMA layer is spin-coated on the SiO_2 film, followed by pre-baking at 180 °C for 3 min. A layer of conductive polymer is then spin-coated on the PMMA to avoid charge accumulation. The PMMA layer is exposed under EBL (ELS-HS50, ELIONIX INC.) for pattern definition. After being immersed in DI water to remove the conductive polymer layer, the patterned sample is developed with methyl isobutyl ketone (MIBK)/ isopropyl alcohol (IPA) for 75 s and is rinsed in IPA for 20 s. An additional Cr layer with 40 nm thickness is deposited on the patterned sample using an E-gun evaporator. Followed by the lift-off process in Acetone, the pattern is transferred into the Cr layer. Taking the Cr layer as the hard mask, the SiO_2 layer is etched by inductively coupled plasma reactive ion etching (ICP-RIE) with CF_4 gas. Chromium etchant is adopted to remove the remaining Cr. A second ICP-RIE with a mixture of Ar and Cl_2 is applied for pattern transfer from the patterned SiO_2 film to the GaN film. After removing the residual SiO_2 using a buffered oxide etch (BOE) solution, the desired GaN nanostructure on the sapphire substrate is finally realized.

Figure 2(a) demonstrates the optical image of fabricated binocular meta-lens. The fabrication process of the well structure was characterized based on scanning electron microscope (SEM) images. There is no cracks or pores on the fabricated nanopillars, as shown in the top-view SEM image of Fig. 2(b). From the zoomed-in tilted view of the nanopillar SEM image in Fig. 2(c), the good collimation of the 750-nm high nanopillars can be observed with precise etching. The physical dimension analysis of the binocular sample is divided into two parts: the sapphire substrate and two GaN meta-lens. Each meta-lens, measuring 2.6 mm in diameter with a volume of 4.25×10^{-6} cm^3, weighs 2.61×10^{-5} g, which is lighter than one percent of the weight of a hair. The weight of the sapphire substrate is 0.115 g and occupies a volume of 0.0288 cm^3. Even though the sapphire substrate brings much more occupation, the overall weight and volume are still tiny and ignorable.

For disparity computation, we propose a pyramid stereo-matching neural network (named H-Net) with a novel "H"-shaped attention module (H-Module), as shown in Fig. 3(a). The H-Net follows an end-to-end learning framework from stereo input images to disparity map prediction without any other pre- or post-processing. The global context aggregation is vital to derive the disparity information from stereo image pairs. Besides the conventional encoder-decoder architecture and pyramid pooling, H-Net adopts cross-pixel interaction and cross-view interaction to enable the utilization of contextual information and the integration of diverse perspectives (see details in Supplementary information Section 4). Compared with the conventional block matching method[43] and two advanced neural networks[34, 44], H-Net demonstrates significant performance improvements and more comprehensive analysis. (see details in Supplementary information Section 5) With the backbone of PSMNet[34], the head of H-Net is a Siamese network[45], whose two branch networks are weight sharing. These head Siamese CNNs utilize residual blocks[46] to extract features and weight-sharing spatial pyramid pooling (SPP) modules[34] to aggregate context information. The output left and right feature maps from the head backbone (Siamese CNNs) are integrated by the proposed H-Module. The introduction of H-Module with attention mechanism[47, 48] allows the model to dynamically allocate its attention or resources based on the relevance or significance of specific features or contexts. H-Module is a symmetric processing pipeline composed

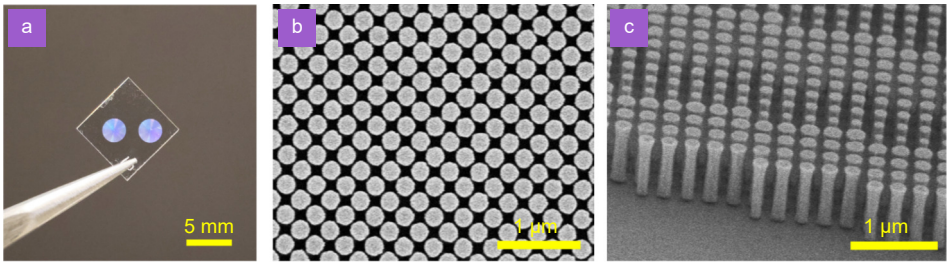

Fig. 2 | Optical and SEM images of fabricated binocular meta-lens. (**a**) Optical image of the binocular meta-lens. (**b**) The zoomed-in top-view SEM image of the meta-lens. (**c**) The zoomed-in tilted-view SEM image at the edge of the meta-lens.

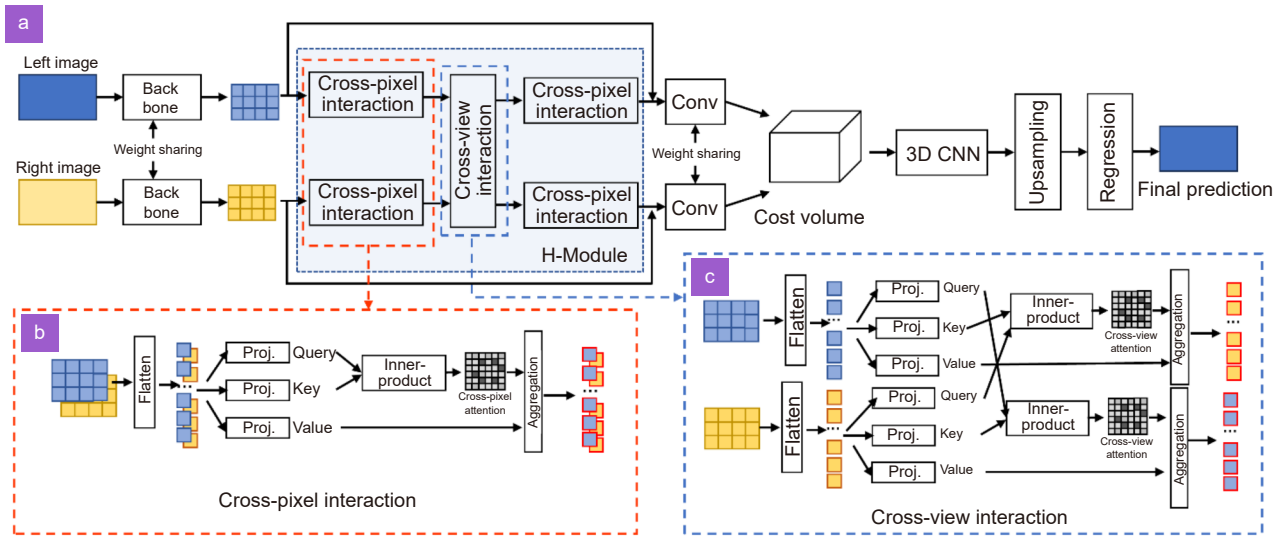

Fig. 3 | Disparity computation with neural network. (**a**) Architecture overview of proposed neural network H-Net with H-Module. The stereo images are processed by weight-sharing backbones to extract features. These features are then combined using cross-pixel interaction and cross-view interaction in an H-Module. A 4D cost volume is created from the left and right image features, which is then used in a 3D CNN for depth estimation. A disparity regression module is performed before the final disparity map prediction. (**b**) Detailed pipelines of the cross-pixel interaction. The left and right feature maps are flattened and processed through separate fully connected layers to generate Query, Key, and Value vectors. The inner product is utilized to compute the similarity between Query and Key, resulting in weight coefficients for each Key. These coefficients are used for cross-pixel attention, associating each Key with its corresponding Value. The weighted Values are aggregated to produce enhanced features. (**c**) Detailed pipelines of the cross-view interaction. The difference from the cross-pixel interaction is the inner product of Key and Query vector comes from different stereo views.

of four cross-pixel interaction blocks and one cross-view interaction block. Cross-pixel interaction is the mutual interaction or influence between pixels in an image or visual representation. It involves considering the relationships and dependencies between neighboring pixels to capture contextual information and improve the understanding or analysis of the image. As illustrated in Fig. 3(b), the left and right feature maps are flattened and projected through separate fully connected layers into three essential vectors: Query, Key, and Value. The similarity or correlation between Query and Key is computed using the inner product, yielding weight coefficients for each Key corresponding to its associated Value, known as cross-pixel attention. The Value is then weighted and aggregated based on attention coefficients to obtain enhanced features. Corresponding attention calculation equation[49] is

$$Attention\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V\,, \tag{2}$$

$$softmax\left(x_i\right)=\frac{e^{x_i}}{\sum_{j=1}^{n}e^{x_j}}\,, \tag{3}$$

where $\boldsymbol{Q}$ is the Query vector, $\boldsymbol{K}$ is the Key vector, $\boldsymbol{V}$ is the Value vector, $\sqrt{d_k}$ serves as a scale to control the result range, d_k is the dimension of Query vector and Key vector, and *softmax* is a normalization function utilized to transform a vector of numerical values into a vector of probability distributions. This transformation ensures that the probability associated with each value is directly proportional to its relative proportion within the original vector.

Cross-view interaction refers to the interaction or integration of information from multiple views or perspectives. In multi-view analysis, cross-view interaction aims to leverage information from different viewpoints or modalities to enhance the overall understanding or interpretation of the scene. Detailed processing steps are depicted in Fig. 3(c), which is similar to cross-pixel interaction. The difference is that the calculation of cross-view attention is based on the Query and Key from different features. Specifically, the Query of the left feature map is computed with the Key of the right feature map through inner product and vice versa. This interaction involves feature matching and data fusion, allowing the alignment and combination of information from different views. The attention mechanism enhances the model's ability to capture dependencies, focus on relevant information, and leverage contextual relationships within the visual data (see the ablation study details in Supplemen-

tary information Section 5.4 Ablation study). The enhanced left and right feature maps from H-Module are concatenated as a 4D cost volume. Three repeated encoder-decoder architectures are utilized in the 3D CNN module to further comprehensively understand the contextual information. Before the final prediction of the disparity map, a disparity regression[50] is performed with a soft attention mechanism. For the disparity map $\mathcal{D} = \{d_a\}_{a=0}^{A_{\max}}$, each final disparity value $\widehat{d}_a$ is the original depth value d_a weighted by its probability. The disparity regression is performed as the equation below

$$\widehat{d}_a = \sum_{a=0}^{A_{\max}} d_a \cdot softmax\left(-c_a\right) \ , \tag{4}$$

where $\widehat{d}_a$ is the final predicted disparity, c_a is the corresponding cost from cost volume for each disparity d_a, a is the annotation number associated with each disparity value d_a in disparity map $\mathcal{D}$, $A_{\max}$ is the maximum value of a within the range of annotations, *softmax* function is discussed in Eq. (3). We adopt the smooth L1 loss as the loss function for its fast convergence and robustness to outliers. The final loss is averaged over the N-pixel disparity map, as shown in Eq. (5).

$$Loss\left(D, \widehat{D}\right) = \frac{1}{N}\sum_{n}^{N} smooth_{L1}\left(d_n - \widehat{d}_n\right) \ , \tag{5}$$

in which

$$smooth_{L1}\left(x\right) = \begin{cases} 0.5x^2, \ \text{if}\ |x| < 1 \\ |x| - 0.5, \ \text{otherwise} \end{cases} \ , \tag{6}$$

where D is the ground truth disparity map, $\widehat{D}$ is the predicted disparity map, N is the number of pixels in the disparity map, d_n is the ground truth disparity data for pixel n, and $\widehat{d}_n$ is the predicted disparity data for pixel n. H-net was trained on the stereo vision dataset KITTI 2012[51]. We employed the Adam Optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$). The learning rate was 0.001 for the first 10 epochs and 0.0001 for the rest. The batch size was 3 on a Nvidia GeForce RTX 3090 GPU. After 800 epochs (64,000 iterations) of training, the final model converged with a training loss of approximately 0.3 (see details in Supporting Information Fig. S8).

The depth map is calculated based on the predicted disparity map. The depth calculation formula[3] is

$$depth = \frac{fb}{ps \cdot \left|\widehat{D} + U_{offs} + O_{offs}\right|} \ , \tag{7}$$

in which

$$O_{offs} = \frac{b}{ps} - \left|x_1 - x_0\right| \ , \tag{8}$$

where focal length f is 10 mm, baseline b is measured 4.056 mm, the side length of the physical pixel on CMOS sensor ps is 3.45 μm, misalignment of lens and sensor on the x-axis U_{offs} is 0, the principal point offset along the x-axis O_{offs} is calculated as -396.6 pixels with x coordinate of left image center $x_0 = 1232$, and the x coordinate of the right image center $x_1 = 2789$ (see more details in Supplementary information Fig. S3). The edge image is derived from the raw captured stereo image with a Canny edge detector[52], which approximates the first derivative of a Gaussian operator. Through the lower bound cut-off suppression and edge tracking by hysteresis, the detected edges are constrained to be one pixel wide and located at the center discontinuous area without false noise edge points. There are no edges in the non-textured regions in images with uniform intensity distribution. These ill-posed regions will cause ambiguous depth prediction because of the feature-matching calculation mechanism. Under the guidance of the edge image, these ill-posed regions on the depth map are discarded. The edge-enhanced depth perception is the depth map filtered by logical conjunction (AND) operations on edge images. Both the discontinuity of intensity and depth are preserved with high fidelity and accuracy.

Results and discussion

The optical performance of the fabricated meta-lens is derived under 532 nm illumination. The measured intensity profile of left and right meta-lenses along the propagation direction is presented in Fig. 4(a). The corresponding measured focal lengths of left and right meta-lenses are 10.048 mm and 10.046 mm, which matches the designed focal length of 10.0 mm. The diameter of a single meta-lens is 2.6 mm, and the metalens' numerical aperture (NA) is about 0.13. The measured full-width at half-maximum (FWHM) of the focal spots of both meta-lens along X- and Y-axes range from 2.21 to 2.36 μm, with the minimum measurement accuracy of 0.2809 μm per division. Therefore, the averaged FWHM is 2.26 ± 0.14 μm, which is close to the diffraction-limited system with an FWHM of 2.1 μm (FWHM = $0.514\lambda/NA$). The modulation transfer function (MTF), the Fourier transform of the point spread function (PSF), was also calculated, which further confirms that the fabricated meta-lens is a diffraction-limited lens (see more

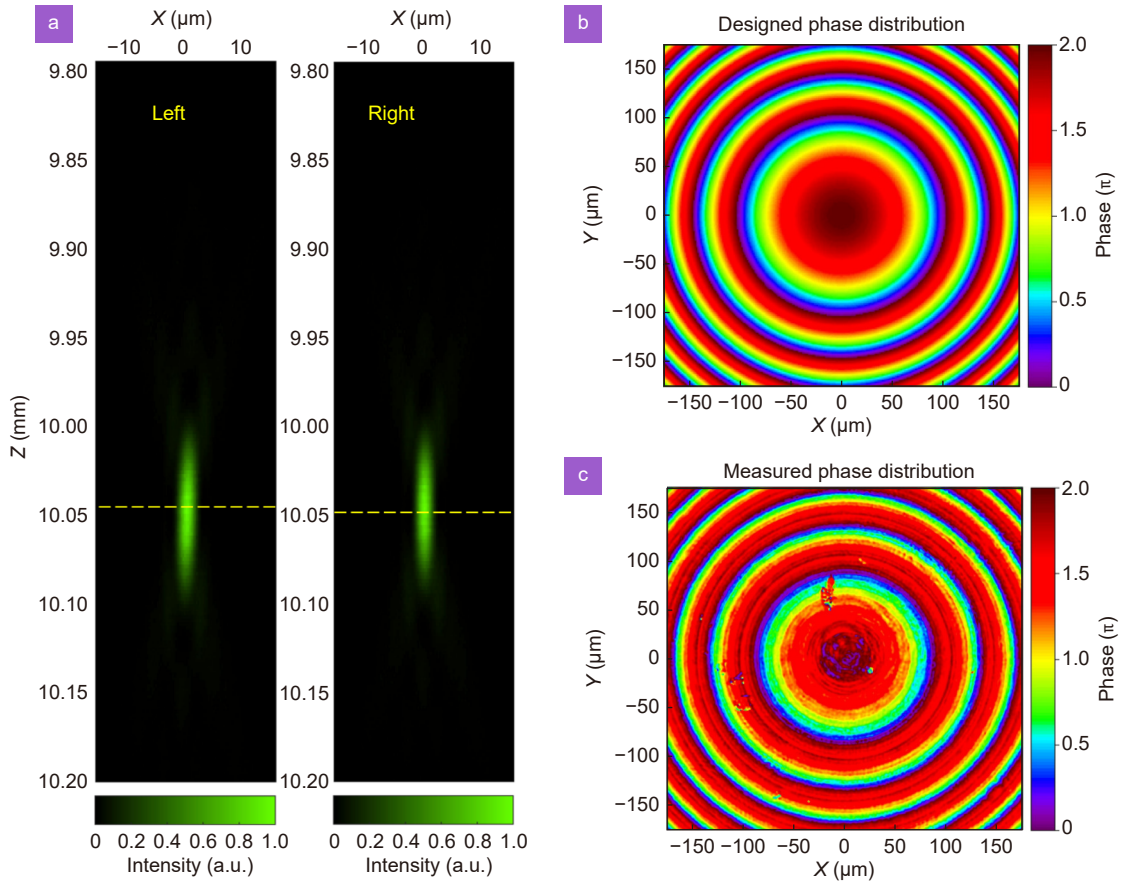

Fig. 4 | Characterization of binocular meta-lens. (**a**) *X*-*Z* plane focusing profiles of left and right meta-lens under 532 nm of wavelength. The measured focal lengths of left and right meta-lenses are 10.048 mm and 10.046 mm, respectively, which are denoted by yellow dashed lines. (**b**) Designed phase distribution of the meta-lens. (**c**) Corresponding measured phase distribution of the meta-lens in (b).

details in Supporting Information Fig. S4). The measured focusing efficiency is 73.86% at the working wavelength of 532 nm. The focusing efficiency is calculated by dividing the total light power of the focal point area at the focal plane by the total input light power of the bare substrate surface (the selected area is equal to the size of the meta-lens). Several experiments were performed to characterize the 2.6 mm meta-lens using a commercial measurement system (AR-Meta-P, IDEAOPTICS INC.). The phase profile of the fabricated meta-lens was measured to check the agreement between the calculated phase profile and the fabricated phase profile. The detailed experimental setup for meta-lens phase measurement was demonstrated in our previous work[53]. The simulated and experimental phase distribution maps at the central region of the meta-lens are depicted in Fig. 4(b) and 4(c), respectively, which are in good agreement with each other. The small disparities can be attributed to the fabrication defects and the spherical aberrations in the measurement system. More theoretical and the measured phase profile comparison results are depicted in Supporting Information Fig. S5.

Various imaging and depth sensing experiments are conducted to test the performance of edge-enhanced depth perception of our binocular meta-lens. The configuration of the binocular meta-lens camera for imaging is shown in Fig. S7 in the Supporting Information. Figure 5 demonstrates the raw captured images, depth sensing results, edge-enhanced depth maps, and the integration results of raw images and 3D edges. The raw captured image I_{raw} is cropped from the common stereoscopic region of the left image. Proposed H-Net outputs corresponding disparity map of the stereo images. Through Eq. (5), the depth map $\widehat{D}_{\mathrm{epth}}$ is calculated accordingly and illustrated in pseudocolor, as shown in the second column of Fig. 5. The 2D edge images that represent the spatial intensity discontinuity are derived from the raw captured image (first column) with the Canny operator. The 2D edge image is converted into a binary matrix E_{b}, in which the edge pixel is 1, otherwise it is 0. The edge-enhanced depth map *DE* is calculated by the Hadamard product of the depth map $\widehat{D}_{\mathrm{epth}}$ and the binary edge ma-

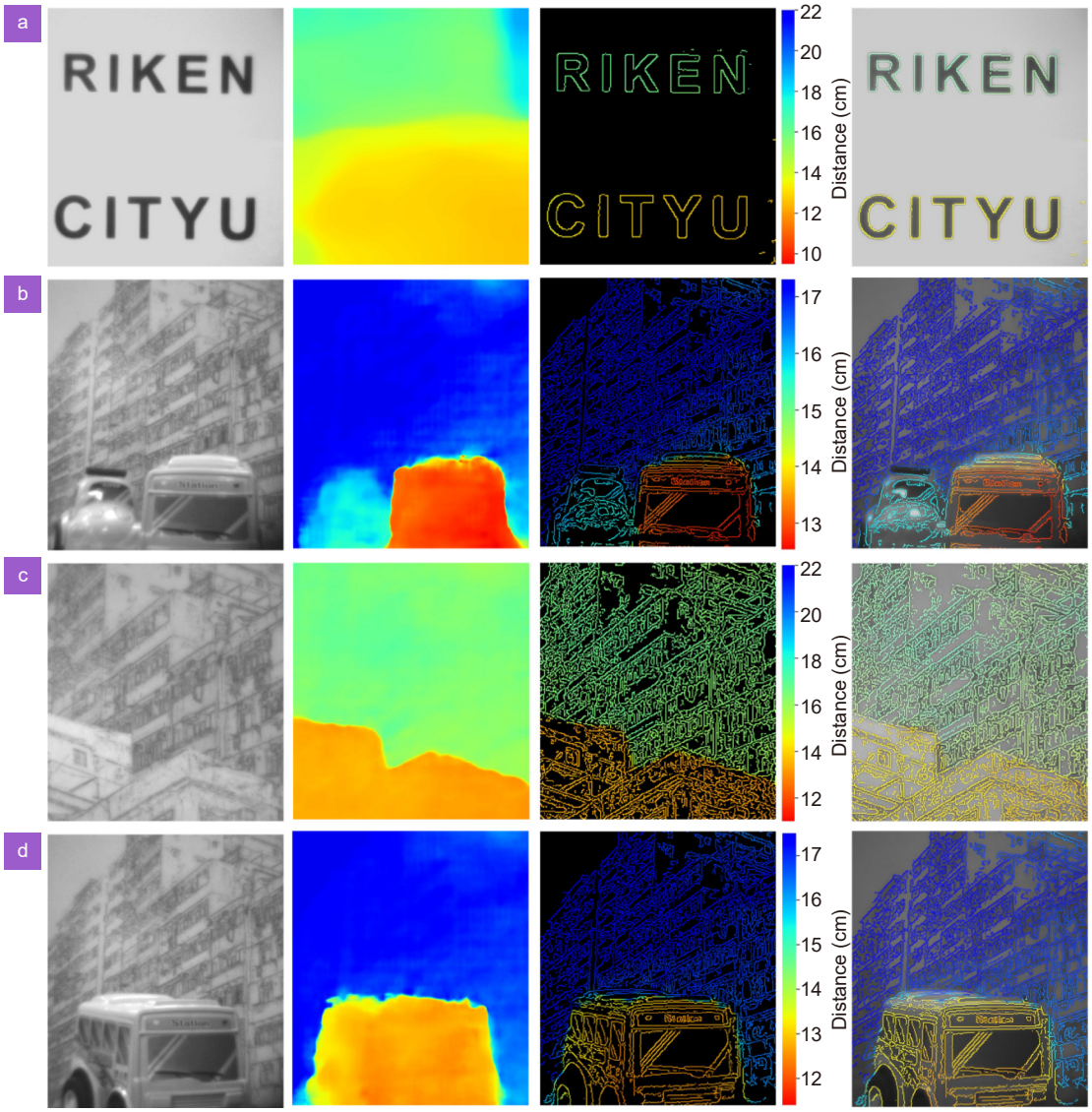

Fig. 5 | Edge-enhanced depth perception of various objects. The first column is the raw left image. The second column is the corresponding depth map. The third column is the edge-enhanced depth map. The second and third columns use the same color bar on the right of the third column. The fourth column is the integration image of the raw image and edge-enhanced depth map. (**a**) Two pieces of transparent plastic paper printed with "RIKEN" and "CITYU" in black letters are placed at 16.0 cm and 12.8 cm, respectively. (**b**) A piece of sketch paper printed with a tilted three-dimensional building is located at 17.3 cm as the background. The front ends of the two toy cars are approximately 12.9 cm and 15.7 cm, respectively. (**c**) The two architectural sketches are at 13.5 cm and 16.5 cm, respectively. (**d**) The background architecture sketch is positioned at 17.3 cm. The depth of a toy car's body ranges from 12.5 cm to 15.5 cm.

trix E_b, which is similar to a logical conjunction (AND) operation. The specific calculation equation is

$$DE = \widehat{D}_{\text{epth}} \odot E_{\text{b}} \ . \tag{9}$$

The edge-enhanced depth maps DE are displayed in the third column of Fig. 5 in pseudocolor. The non-edge regions with 0 values are set to be black. The integration images I_{integ} in the fourth column of Fig. 5 are merged using the following expression:

$$I_{\text{integ}} = 1.2DE + 0.8I_{\text{raw}} \ . \tag{10}$$

The integration images aim to demonstrate the fidelity of edge-enhanced depth perception in spatial intensity and depth discontinuity detection.

Figure 5(a) depicts a scenario with ill-posed regions. Two black letter objects, "RIKEN" and "CITYU," printed on transparent plastic papers, are positioned at 16.0 cm and 12.8 cm, respectively. The letter carrier is transparent plastic paper. The background is a white wall without any texture. The absence of texture makes it difficult to establish reliable correspondences between image points in the left and right views, leading to unreliable or

ambiguous depth estimates (see the middle region of the depth map in Fig. 5(a). Such unreliable and ambiguous depth estimates will cause severe trouble for decision-making tasks. In edge-enhanced depth perception, the 3D edge data agree well with the ground truth with the completed preservation of essential details of the scene. Figure 5(b) demonstrates a multi-object traffic scene with two toy cars located at about 12.9 cm and 15.7 cm. An architecture sketch background providing is placed at 17.3 cm. Figure 5(c) shows two architecture sketches with false 3D feelings positioned at 13.5 cm and 16.5 cm, respectively. With edge-enhanced depth perception, the planar false 3D objects do not deceive the system. Figure 5(d) displays a toy car with a continuous depth change, ranging from 12.5 cm to 15.5 cm. All depth sensing results are correct, demonstrating the accuracy capability of our H-Net. The edge-enhanced depth results discard all uniform regions and amplify the 3D feature details with high confidence.

Conclusions

Spatial computing has attracted growing attention from both academia and industry, driven by the rising popularity of the metaverse. A portable and accurate imaging and depth sensing system is of vital importance for next-generation virtual reality devices. In this work, we demonstrate an edge-enhanced depth perception system based on binocular meta-lens, which is intelligent, lightweight, and compact for spatial computing. The miniaturized system contains a binocular meta-lens, a 532 nm filter, and a CMOS sensor. The binocular meta-lens only weighs about 0.115 g with 0.0288 cm^3 volume consumption. The imaging system based on our meta-lens minimizes the discomfort caused by the weight and volume of wearable devices to users. We propose a stereo-matching neural network with a novel H-Module for the disparity computation. The H-Module introduces the attention mechanism into the Siamese network. The symmetric architecture with cross-pixel interaction and cross-view interaction enables a more comprehensive analysis of the contextual information in stereo images. The edge enhancement based on the spatial intensity discontinuity discards the ill-posed regions in the image, where ambiguous depth prediction will be generated due to the lack of texture information. With the support of deep learning, our edge-enhanced provides a prompt, intelligent response in less than 0.15 seconds. This edge-enhanced depth perception system will facilitate accurate 3D scene modeling to promote the development of machine vision, autonomous driving, and robotics.

References

1. Greenwold S. Spatial computing (Massachusetts Institute of Technology, Cambridge, 2003).
2. Pangilinan E, Lukas S, Mohan V. *Creating Augmented and Virtual Realities: Theory and Practice for Next-Generation Spatial Computing* (O'Reilly Media, Inc. , Sebastopol, 2019).
3. Liu XY, Chen MK, Chu CH et al. Underwater binocular meta-lens. *ACS Photonics* **10**, 2382–2389 (2023).
4. Chen MK, Chu CH, Liu XY et al. Meta-lens in the sky. *IEEE Access* **10**, 46552–46557 (2022).
5. Jeon D, Shin K, Moon SW et al. Recent advancements of metalenses for functional imaging. *Nano Convergence* **10**, 24 (2023).
6. Li T, Chen C, Xiao XJ et al. Revolutionary meta-imaging: from superlens to metalens. *Photon Insights* **2**, R01 (2023).
7. Moon SW, Lee C, Yang Y et al. Tutorial on metalenses for advanced flat optics: design, fabrication, and critical considerations. *J Appl Phys* **131**, 091101 (2022).
8. Pu MB, Li X, Ma XL et al. Catenary optics for achromatic generation of perfect optical angular momentum. *Sci Adv* **1**, e1500396 (2015).
9. Hu YQ, Li L, Wang YJ et al. Trichromatic and tripolarization-channel holography with noninterleaved dielectric metasurface. *Nano Lett* **20**, 994–1002 (2020).
10. Wu PC, Sokhoyan R, Shirmanesh GK et al. Near - infrared active metasurface for dynamic polarization conversion. *Adv Opt Mater* **9**, 2100230 (2021).
11. Song QH, Baroni A, Wu PC et al. Broadband decoupling of intensity and polarization with vectorial Fourier metasurfaces. *Nat. Commun* **12**, 3631 (2021).
12. Song MW, Feng L, Huo PC et al. Versatile full-colour nanopainting enabled by a pixelated plasmonic metasurface. *Nat Nanotechnol* **18**, 71–78 (2023).
13. Li X, Chen QM, Zhang X et al. Time-sequential color code division multiplexing holographic display with metasurface. *Opto-Electron Adv* **6**, 220060 (2023).
14. Guo YH, Pu MB, Zhang F et al. Classical and generalized geometric phase in electromagnetic metasurfaces. *Photon Insights* **1**, R03 (2022).
15. Wang SM, Wu PC, Su VC et al. A broadband achromatic metalens in the visible. *Nat Nanotechnol* **13**, 227–232 (2018).
16. Zhang F, Pu MB, Li X et al. Extreme-angle silicon infrared optics enabled by streamlined surfaces. *Adv Mater* **33**, 2008157 (2021).
17. Fan QB, Xu WZ, Hu XM et al. Trilobite-inspired neural nanophotonic light-field camera with extreme depth-of-field. *Nat Commun* **13**, 2130 (2022).
18. Chen MK, Liu XY, Sun YN et al. Artificial Intelligence in Meta-optics. *Chem Rev* **122**, 15356–15413 (2022).
19. Krasikov S, Tranter A, Bogdanov A et al. Intelligent metaphotonics empowered by machine learning. *Opto-Electron Adv* **5**, 210147 (2022).
20. Chen MK, Liu XY, Wu YF et al. A meta-device for intelligent depth perception. *Adv Mater* **35**, 2107465 (2023).
21. Li ZS, Sun JS, Fan Y et al. Deep learning assisted variational Hilbert quantitative phase imaging. *Opto-Electron Sci* **2**, 220023 (2023).
22. Liu C, Ma Q, Luo ZJ et al. A programmable diffractive deep neural network based on a digital-coding metasurface array. *Nat*

Electron **5**, 113–122 (2022).
23. Gao XX, Ma Q, Gu Z et al. Programmable surface plasmonic neural networks for microwave detection and processing. *Nat Electron* **6**, 319–328 (2023).
24. Li LL, Ruan HX, Liu C et al. Machine-learning reprogrammable metasurface imager. *Nat Commun* **10**, 1082 (2019).
25. Li LL, Zhao HT, Liu C et al. Intelligent metasurfaces: control, communication and computing. *eLight* **2**, 7 (2022).
26. Blake R, Wilson H. Binocular vision. *Vision Res* **51**, 754–770 (2011).
27. Hirschmuller H. Accurate and efficient stereo processing by semi-global matching and mutual information. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* 807–814 (IEEE, 2005); http://doi.org/10.1109/CVPR.2005.56.
28. Zhou HQ, Wang YT, Li X et al. A deep learning approach for trustworthy high-fidelity computational holographic orbital angular momentum communication. *Appl Phys Lett* **119**, 044104 (2021).
29. He C, Zhao D, Fan F et al. Pluggable multitask diffractive neural networks based on cascaded metasurfaces. *Opto-Electron Adv* **7**, 230005 (2024).
30. Liao MH, Zheng SS, Pan SX et al. Deep-learning-based ciphertext-only attack on optical double random phase encryption. *Opto-Electron Adv* **4**, 200016 (2021).
31. Hao JY, Lin X, Lin YK et al. Lensless complex amplitude demodulation based on deep learning in holographic data storage. *Opto-Electron Adv* **6**, 220157 (2023).
32. Ma TG, Tobah M, Wang HZ et al. Benchmarking deep learning-based models on nanophotonic inverse design problems. *Opto-Electron Sci* **1**, 210012 (2022).
33. Lin CH, Huang SH, Lin TH et al. Metasurface-empowered snapshot hyperspectral imaging with convex/deep (CODE) small-data learning theory. *Nat Commun* **14**, 6979 (2023).
34. Chang JR, Chen YS. Pyramid stereo matching network. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 5410–5418 (IEEE, 2018); http://doi.org/10.1109/CVPR.2018.00567.
35. Zhou Y, Zheng HY, Kravchenko II et al. Flat optics for image differentiation. *Nat Photonics* **14**, 316–323 (2020).
36. Guo C, Xiao M, Minkov M et al. Photonic crystal slab Laplace operator for image differentiation. *Optica* **5**, 251–256 (2018).
37. Kim Y, Lee GY, Sung J et al. Spiral metalens for phase contrast imaging. *Adv Funct Mater* **32**, 2106050 (2022).
38. Chen MK, Yan Y, Liu XY et al. Edge detection with meta-lens: from one dimension to three dimensions. *Nanophotonics* **10**, 3709–3715 (2021).
39. Badloe T, Kim Y, Kim J et al. Bright-field and edge-enhanced imaging using an electrically tunable dual-mode metalens. *ACS Nano* **17**, 14678–14685 (2023).
40. Huo PC, Zhang C, Zhu WQ et al. Photonic spin-multiplexing metasurface for switchable spiral phase contrast imaging. *Nano Lett* **20**, 2791–2798 (2020).
41. Zhou JX, Qian HL, Chen CF et al. Optical edge detection based on high-efficiency dielectric metasurface. *Proc Natl Acad Sci USA* **116**, 11137–11140 (2019).
42. Zhou JX, Liu SK, Qian HL et al. Metasurface enabled quantum edge detection. *Sci Adv* **6**, eabc4385 (2020).
43. Hamid MS, Manap NA, Hamzah RA et al. Stereo matching algorithm based on deep learning: A survey. *J King Saud Univ - Comput Inf Sci* **34**, 1663–1673 (2022).
44. Xu HF, Zhang J, Cai JF et al. Unifying flow, stereo and depth estimation. *IEEE Trans Pattern Anal Mach Intell* **45**, 13941–13958 (2023).
45. Taigman Y, Yang M, Ranzato MA et al. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition* 1701–1708 (IEEE, 2014); http://doi.org/10.1109/CVPR.2014.220.
46. He KM, Zhang XY, Ren SQ et al. Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (IEEE, 2016); http://doi.org/10.1109/CVPR.2016.90.
47. Li WY, Liu XY, Yuan YX. SIGMA: Semantic-complete graph matching for domain adaptive object detection. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 5281–5290 (IEEE, 2022); http://doi.org/10.1109/CVPR52688.2022.00522.
48. Li WY, Liu XY, Yuan YX. SIGMA++: Improved semantic-complete graph matching for domain adaptive object detection. *IEEE Trans Pattern Anal Mach Intell* **45**, 9022–9040 (2023).
49. Vaswani A, Shazeer N, Parmar N et al. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* 6000–6010 (Curran Associates Inc. , 2017).
50. Kendall A, Martirosyan H, Dasgupta S et al. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the 2017 IEEE International Conference on Computer Vision* 66–75 (IEEE, 2017); http://doi.org/10.1109/ICCV.2017.17.
51. Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* 3354–3361 (IEEE, 2012); http://doi.org/10.1109/CVPR.2012.6248074.
52. Canny J. A computational approach to edge detection. *IEEE Trans Pattern Anal Mach Intell* **PAMI-8**, 679–698 (1986).
53. Zhao MX, Chen MK, Zhuang ZP et al. Phase characterisation of metalenses. *Light Sci Appl* **10**, 52 (2021).

Acknowledgements

We are grateful for financial supports from the Research Grants Council of the Hong Kong Special Administrative Region, China [Project No. C5031-22G; CityU11310522; CityU11300123], the Department of Science and Technology of Guangdong Province [Project No. 2020B1515120073], City University of Hong Kong [Project No. 9610628] and, JST CREST (Grant No. JPMJCR1904) .

Author contributions

XY Liu, T Tanaka, and MK Chen. organized the project. XY Liu, BR Leng, JC Zhang, Y Zhou, JL Cheng, and MK Chen conceived the principle, numerical design, and characterization of the metasurface and meta-system. T Yamaguchi and T Tanaka conceived the fabrication of metasurface. XY Liu, and MK Chen built up the optical system for measurement and the deep learning model and collected the data with experiments for analysis. All authors discussed the results, prepared the manuscripts, and commented on the manuscript.

Competing interests

The authors declare no competing financial interests.

Supplementary information

Supplementary information for this paper is available at https://doi.org/10.29026/oes.2024.230033

Opto-Electronic Science

Supplementary information

2024, Vol. 3, No.

DOI: 10.29026/oes.2024.230033

Edge enhanced depth perception with binocular meta-lens

Xiaoyuan Liu[1,2,3], Jingcheng Zhang[1], Borui Leng[1], Yin Zhou[1], Jialuo Cheng[1], Takeshi Yamaguchi[4,5,6], Takuo Tanaka[4,5,6]* and Mu Ku Chen[1,2,3]*

[1]Department of Electrical Engineering, City University of Hong Kong, Hong Kong SAR 999077, China; [2]Centre for Biosystems, Neuroscience, and Nanotechnology, City University of Hong Kong, Hong Kong SAR 999077, China; [3]The State Key Laboratory of Terahertz and Millimeter Waves, and Nanotechnology, City University of Hong Kong, Hong Kong SAR 999077, China; [4]Innovative Photon Manipulation Research Team, RIKEN Center for Advanced Photonics, 351-0198, Japan; [5]Metamaterial Laboratory, RIKEN Cluster for Pioneering Research, 351-0198, Japan; [6]Institute of Post-LED Photonics, Tokushima University, 770-8506, Japan.

*Correspondence: T Tanaka, E-mail: t-tanaka@riken.jp; MK Chen, E-mail: mkchen@cityu.edu.hk

This file includes:

Supplementary information for this paper is available at https://doi.org/10.29026/oes.2024.230033

Section 1: Design and fabrication of the binocular meta-lens

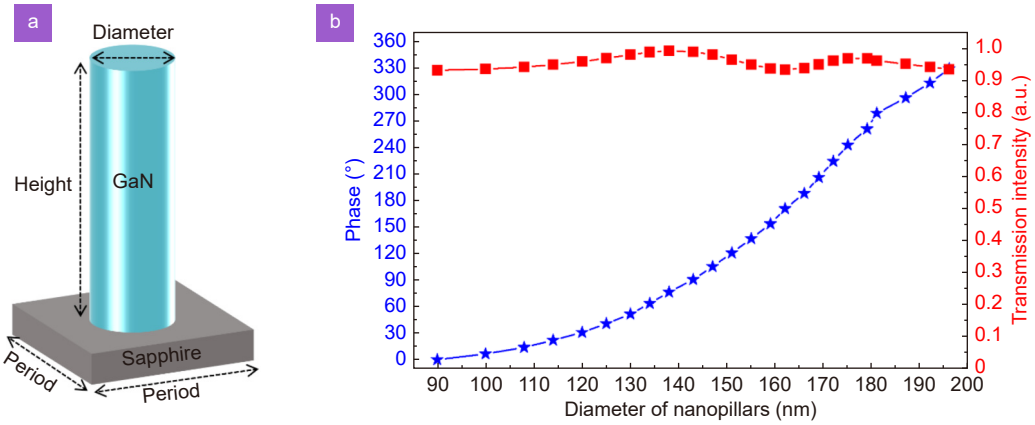

Fig. S1 | Meta-atom design and simulation of the binocular meta-lens. (**a**) The Schematic diagram of the GaN meta-lens fabrication process. (**b**) The phase modulation and transmission intensity of the meta-atom with various nanopillar diameters.

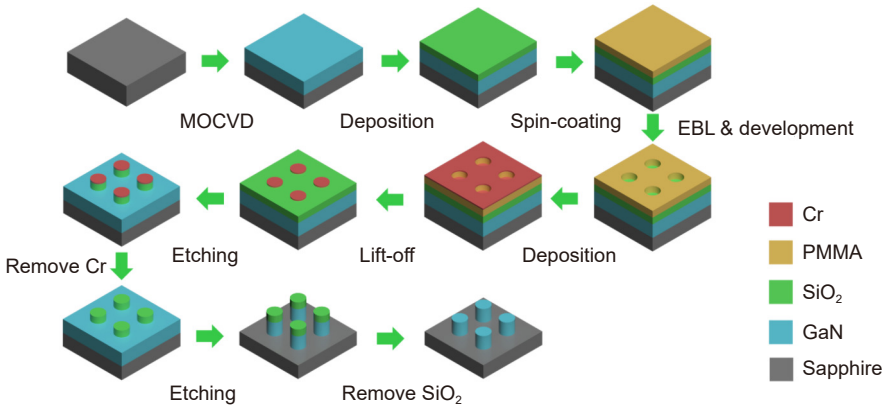

Fig. S2 | The Schematic diagram of the GaN meta-lens fabrication process.

Section 2: Characterization of binocular meta-lens

The depth resolution and accuracy are related to the object depth itself, as shown in Fig. S3(b). The closer the object is to the meta-lens, the higher the depth resolution and accuracy will be. In the expected distance range to be measured, the smaller the slope of the data curve is, the higher the spatial resolution is. For example, for a distance below 100 mm, if the distance changes slightly, the disparity is changed significantly. The yellow curve line is the theoretical value, and the violet point is the experimental value. The measured results agree well with the theoretical results.

The highest accuracy of our meta-lens system is determined by Eq. (S1).

$$acc = \frac{fb}{ps}\left(\frac{1}{O_{\text{offs}} - 1} - \frac{1}{O_{\text{offs}}}\right), \tag{S1}$$

where focal length f is 10 mm, baseline b is measured 4.056 mm, the side length of the physical pixel on CMOS sensor ps is 3.45 μm, the principal point offset along the x-axis O_{offs} is calculated as -396.6 pixels for the experiment demonstration depth working range. Under this configuration, the highest accuracy can reach 74.5 um.

For the working range of 60 to 450 mm in the experimental demonstration of our work, we did a series of scanning measurement experiments to evaluate its depth resolution. A textured pattern was attached to the surface of a flat board. The flat board moved from a distance of 60 mm to 450 mm in 10 mm steps. The distance refers to the separation length between the binocular meta-lens and the flat board. We captured images every time the flatboard moved. We did 10 groups of such scanning measurements for statistical analysis. The measurement results, as depicted in Fig. S4, demonstrate strong agreement between the measured distances and the corresponding ground truth values. Fig. S4(a) showcases the excellent alignment between the measured distances and ground truths, with minimal error bars indicating the

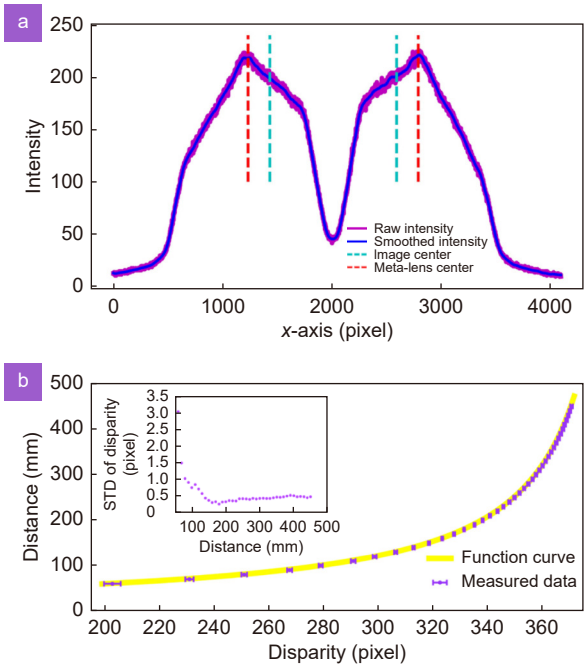

Fig. S3 | Depth calculation analysis based on our binocular meta-lens. (**a**) Intensity distribution along the cut line that crosses the two image centers when photographing a large white object. The distance between the red and cyan dashed lines is 0.5 O_{offs}. (**b**) The function relationship between disparity and distance with experimental verification. The inserted image is the STD distribution of disparity.

absence of crosstalk between measurements. Notably, both the negative error bars in Fig. S4(b) and positive error bars in Fig. S4(c) generally remain below 1 mm. The presence of two outliers can be attributed primarily to errors within the measurement system. As a result, we can confidently conclude that a depth resolution of 1 mm can be reliably achieved within the range of 60 to 450 mm.

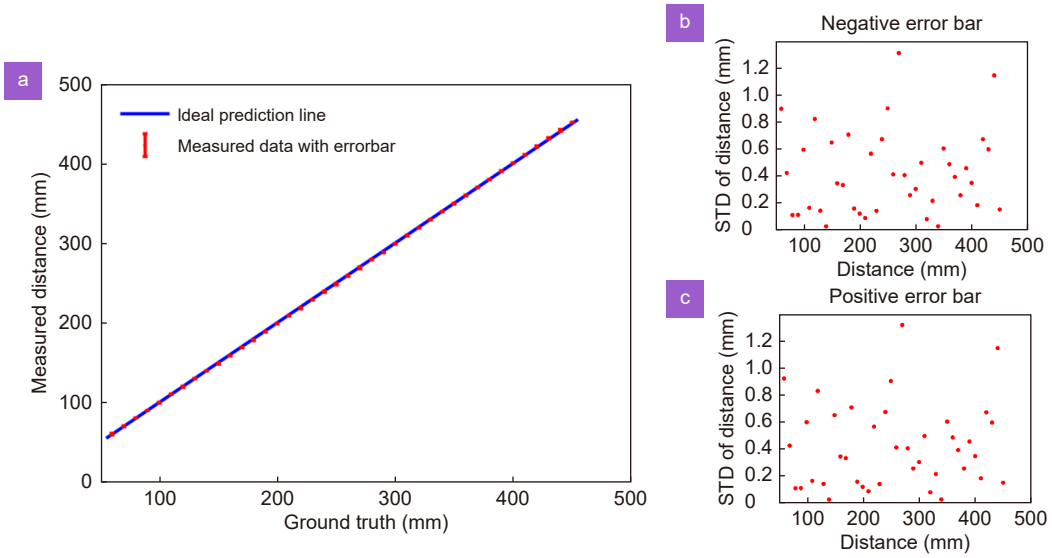

Fig. S4 | Depth resolution in the working range of 60 to 450 mm. (**a**) The measured distance with errorbar versus the ground truth distance. The ideal prediction line, depicted in blue, represents perfect agreement between measurements and ground truth values. The red dots represent the mean values of the measured distances obtained from ten series of scanning experiments, while the error bars illustrate the standard deviation (STD) calculated from these ten groups of scanning measurements. The length of the error is calculated from the standard deviation (STD) of the 10 groups of scanning measurements. The error bars are very small, which are further illustrated in (**b**) and (**c**). (b) The length distribution of the negative error bars in relation to the distances. (b) The length distribution of the positive error bars in relation to the distances.

Due to the limitation of our experimental room size, we discuss the depth resolution at longer working distances through computation.

$$depth = \frac{fb}{ps \cdot \left| \widehat{D} + U_{\text{offs}} + O_{\text{offs}} \right|} , \tag{S2}$$

In the depth calculation Eq. (S2), O_{offs} in our system is 0, $O_{\text{diff}} < 0$ and $\widehat{D} < |O_{\text{diff}}|$. Therefore, Eq. (S2) could be simplified as

$$depth = -\frac{fb}{ps * \left(\widehat{D} + O_{\text{offs}} \right)} , \tag{S3}$$

The depth resolution is related to the object's depth itself. The closer the object is, the higher the depth resolution of the system is. The uncertainty of the depth perception $\Delta depth$ is related to the disparity vibration $\Delta disp$. The disparity vibration $\Delta disp$ is determined by the disparity computation algorithm and the texture of the object. Normally, the disparity vibration $\Delta disp$ is at the subpixel level because the disparity computation algorithms will take global context characteristics into account.

$$\Delta depth = -\frac{fb}{ps} \left(\frac{1}{\widehat{D} + \Delta disp + O_{offs}} - \frac{1}{\widehat{D} + O_{offs}} \right) , \tag{S4}$$

According to Eq. (S3), $\widehat{D}$ could be expressed as

$$\widehat{D} = -\frac{fb}{ps * depth} - O_{offs} , \tag{S5}$$

Putting Eq. (S5) into Eq. (S4), we can derive the depth resolution at different depths,

$$\Delta depth = \frac{A * depth^2}{1 - A * depth}, where A = \frac{ps * \Delta disp}{fb} , \tag{S6}$$

The above discussion is based on the object distance being large (far to meta-lens). In other words, the distance between the meta-lens and sensor could be approximated as focal length. In practical applications, the design parameters of binocular meta-lens, namely the focal length f and baseline b, can be adjusted based on the actual working distance, range, and required accuracy. The focal length f and the baseline b play vital roles in determining the depth sensing accuracy.

The spatial resolution of the lens is usually described by Modulation Transfer Function (MTF). It quantifies the ability of a lens system to transmit details at different spatial frequencies, i.e., how many image details a lens can retain and reproduce. The modulation is typically measured by imaging the object of periodic bright and dark line pairs. The specific calculation of modulation is defined as

$$M = \frac{I_{max} - I_{min}}{I_{max} + I_{min}} , \tag{S7}$$

where I_{max} is the maximum intensity value in the captured image, representing the bright (white) line; I_{min} is the minimum intensity value in the captured image, representing the dark (black) line. MTF reflects the image contrast over different spatial frequencies. Spatial frequency can be described by the number of line pair periods contained within one millimeter in the image. The number of cycles contained in each millimeter on the image plane is called the spatial frequency. Fig. S5(e) demonstrates the measured MTF of our binocular meta-lens. The black dashed line in Fig. S5(e) is the diffraction-limited transfer function. The diffraction limit represents the spatial resolution of the ideal image. The MTF will decrease as the spatial resolution increases. The diffraction limit is calculated as shown in Eq. (S8-S10).

$$MTF(\xi) = \frac{2}{\pi} \left(\phi - \cos\phi \cdot \sin\phi \right) , \tag{S8}$$

$$\phi = \arccos\left(\frac{\xi}{\xi_c} \right) , \tag{S9}$$

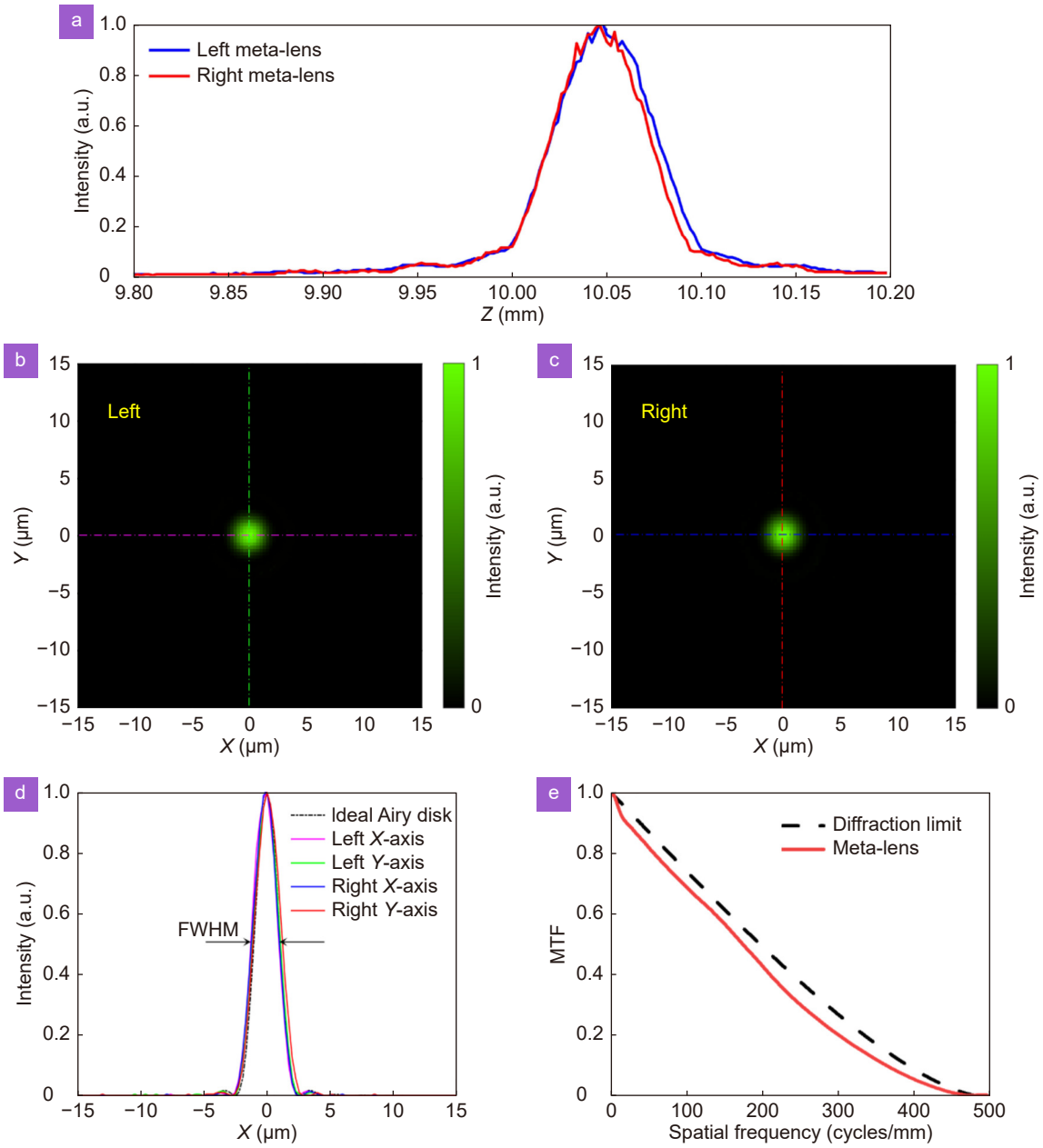

Fig. S5 | The optical performance of GaN meta-lens under 532 nm laser illumination. (**a**) The measured intensity profiles of left and right meta-lens along the optical axes (*z*-axis). (**b**) The measured focal spot image of the left meta-lens at the focal plane (*z* = 10.048 mm). (**c**) The measured focal spot image of the right meta-lens at the focal plane (*z* = 10.046 mm). (**d**) The ideal and measured cross-section intensity profiles of focal spot for both meta-lenses. The measured intensity lines are cut along the *x* and *y* axes (denoted in (b) and (c)), with the brightest point at the focus as the center. The measured FWHM of the focal spots are 2.229 μm of left meta-lens along the *x*-axis, 2.214 μm of left meta-lens along the *y*-axis, 2.231 μm of right meta-lens along the *x*-axis, 2.356 μm of right meta-lens along the *y*-axis, respectively. (**e**) The MTFs of our meta-lens and ideal lens. The red solid line is the measured modulation (contrast) of our meta-lens. The black dashed line is the diffraction limit, which illustrates the theoretical performance expected from an ideal perfect lens.

$$\xi_c = \frac{1}{\lambda \cdot N}, \tag{S10}$$

where ξ is the spatial frequency, ξ_c is the limit frequency (MTF cut-off), λ is the working wavelength of the incident light, N is the f-number given by $N = \frac{f}{D}$. For our binocular meta-lens with focal length $f = 10$ mm and diameter $D = 2.6$ mm, the f-number is 3.846. Corresponding limit spatial frequency ξ_c is 489 cycles/mm under the working wavelength of 532 nm. The measured modulation transfer function (MTF) of our meta-lens (represented by the red

solid line in Fig. S5(e)) closely approximates the diffraction limit (indicated by the black dashed line in Fig. S5(e)). This suggests that the spatial resolution of our meta-lens approaches the level of ideal image quality. Our meta-lens exhibits a notable capability of delivering high image contrast across a wide range of spatial frequencies.

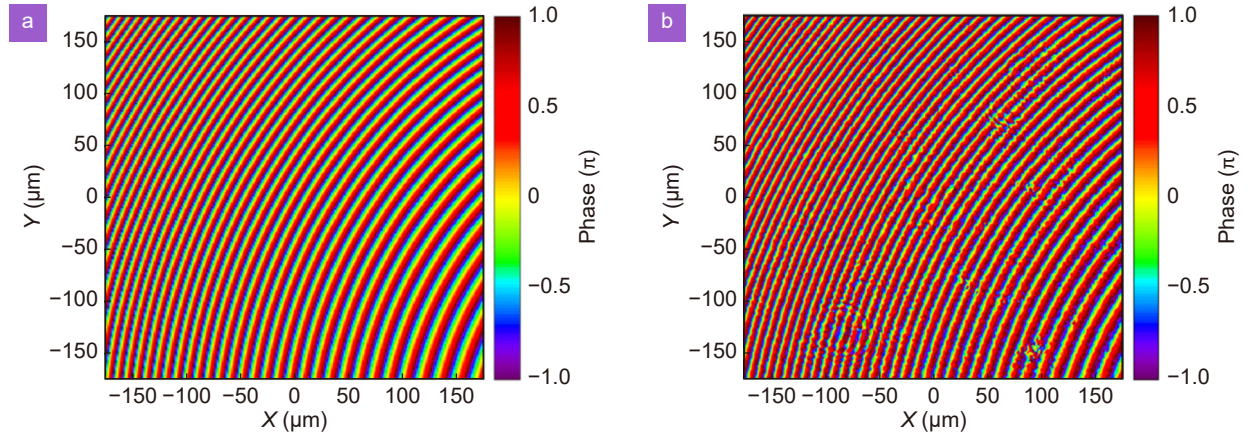

Fig. S6 | The phase characterization of GaN meta-lens. The calculated (**a**) and measured (**b**) phase profile at the edge region of the meta-lens.

Section 3: Configuration of the binocular meta-lens camera

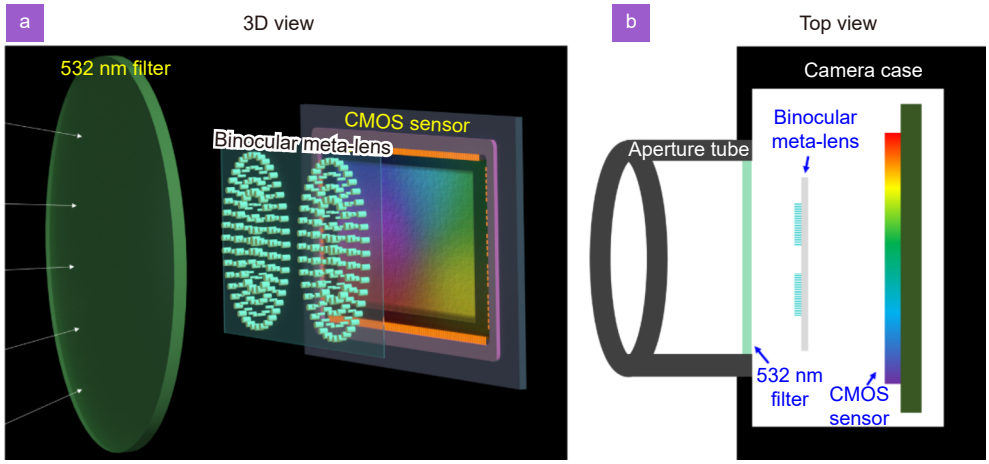

Fig. S7 | The configuration of the binocular meta-lens camera. The 3D view (**a**) and top view (**b**) of the binocular meta-lens camera.

Section 4: Cross-pixel and cross-view interactions

Cross-pixel interaction, also known as spatial interaction, is the mutual interaction or influence between neighboring pixels in an image or visual representation. It involves considering the relationships and dependencies between pixels to capture contextual information and improve the understanding or analysis of the image. Cross-pixel interactions are important for computer vision tasks, such as stereo matching, which strongly relies on image features. The convolution operation in convolutional neural networks (CNN) is a kind of typical cross-pixel interaction.[S1] CNN applies kernels to local patches of the input image. The convolutional operation can be represented mathematically as follows:

$$C(x,y) = k * I(x,y) = \sum_{i=-l}^{l}\sum_{j=-l}^{l} k(i,j)\, I(x-i, y-j) \ , \tag{S11}$$

where $C(x,y)$ represents the convolution output at the position (x,y), k is the kernel with a dimension of $(2l+1, 2l+1)$, $k(i,j)$ represents the kernel value at position (i,j), $I(x-i, y-j)$ represents the input image pixel at the relative position $(x-i, y-j)$. This operation allows the network to learn spatial patterns and dependencies between neighboring pixels. However, the receptive field $\mathcal{I} = \{I_{ij}\}_{i,j=1}^{2l+1}$ in CNN is limited by the kernel size $2l+1$[S2]. Traditional CNNs capture local

relationships through convolutional kernels, but they may struggle to model long-range dependencies between distant pixels in an image.

One of the key challenges in stereo matching is dealing with the ill-posed regions caused by the presence of textureless or repetitive regions in the images. The convolution operation yields local features from small image patches in local neighborhood $\sum_{i=-l}^{l}\sum_{j=-l}^{l} I(x-i, y-j)$, facilitating the establishment of initial feature maps. However, in scenarios where textureless or repetitive regions are present, a broader context $\mathcal{I} = \{I_{ij}\}_{i,j=1}^{P}$, where $P \gg 2l+1$, is necessary, and thus global features come into play. In stereo matching, the extraction of global features entails capturing dependencies between pixels that may not be spatially adjacent. To incorporate contextual information and enable global feature extraction, we introduce the self-attention mechanism[S3] within the cross-pixel interaction module.

Self-attention provides a solution to this problem by allowing each pixel to attend to other pixels in the image, which may not be spatially neighbored. In specific operation, we flattened the $M \times N$ feature map to a sequence of pixels $\mathcal{P} = \{p_i\}_{i=1}^{M\times N}$, where M and N are the height and width of the feature map. For each pixel p_i, we project it into three essential vectors, Query $\boldsymbol{Q}$, Key $\boldsymbol{K}$, and Value $\boldsymbol{V}$, through respective fully connected layers. These linear transformations from fully connected layers map the original pixel representations into higher-dimensional spaces, allowing the model to capture complex cross-pixel relationships and potential contextual information. Self-attention enables the cross-pixel interaction module to compute a weighted sum of the pixel representations, where the weights are determined based on the relevancy or importance of each pixel to the others. Corresponding attention calculation equations[S3] are

$$Attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\boldsymbol{V}, \tag{S12}$$

$$softmax(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}, \tag{S13}$$

where Q is the Query vector, K is the Key vector, V is the Value vector, $\sqrt{d_k}$ serves as a scale to control the result range, d_k is the dimension of the Query vector and Key vector, and *softmax* is a normalization function utilized to transform a vector of numerical values into a vector of probability distributions. The similarity or correlation between Query and Key is computed using the inner product, yielding weight coefficients for each Key corresponding to its associated Value, known as cross-pixel attention. The dot products are then scaled by a factor of the square root of the dimension $\sqrt{d_k}$ to prevent large values. The resulting dot products are passed through a *softmax* function to obtain the cross-pixel attention, which indicates the importance of each pixel for the given Query. This *softmax* transformation ensures that the probability associated with each value is directly proportional to its relative proportion within the original vector. The Value is then weighted and aggregated based on cross-pixel attention to obtain enhanced features. This weighted sum represents the cross-pixel interaction or the aggregated information from the other pixels. The output of the self-attention mechanism for each pixel is a new representation that combines information from both local and distant pixels, allowing the model to capture long-range dependencies. Self-attention enables cross-pixel interactions by allowing each pixel to attend to other pixels in the image. By calculating cross-pixel attention weights based on the relevancy of each pixel, the model can aggregate information from all pixels to generate a new representation that captures both local and long-range dependencies.

In stereo matching, the goal is to determine the correspondence between pixels in a pair of stereo images, which allows for the estimation of disparity information. Therefore, a strong relationship and correspondence exist between the pixels in the left view, denoted as $\mathcal{P}_l = \{p_{l_i}\}_{i=1}^{M\times N}$, and the right view, denoted as $\mathcal{P}_r = \{p_{r_i}\}_{i=1}^{M\times N}$. Cross-view interaction refers to the process of integrating or exchanging information between the left and right stereo views. In binocular-view analysis, our cross-view interaction aims to leverage information from stereo viewpoints or modalities to enhance the overall understanding or interpretation of the scene. Detailed processing steps are similar to the cross-pixel interaction. The distinction lies in the calculation of cross-view attention, which is based on the Query and Key derived from different views. Specifically, the Query of the left feature map Q_l is computed with the Key of the right feature map K_r through inner product and vice versa, as described in Eq. (S14) and (S15).

$$Attention\_left\left(Q_r, K_l, V_l\right) = softmax\left(\frac{Q_r K_l^T}{\sqrt{d_k}}\right) V_l \,, \tag{S14}$$

$$Attention\_right\left(Q_l, K_r, V_r\right) = softmax\left(\frac{Q_l K_r^T}{\sqrt{d_k}}\right) V_r \,, \tag{S15}$$

where Q_l, K_l, V_l are the three essential vectors, Query, Key, and Value, projected from the left feature map; Q_r, K_r, V_r are the three essential vectors, Query, Key, and Value, projected from the right feature map. The inner products of Query and Key vectors from different views indicate the significance or correspondence of each pixel in the current view regarding the given Query from the other view. This cross-view interaction involves feature matching and data fusion, allowing the alignment and combination of information from stereo views. The cross-attention mechanism enhances the model's ability to capture dependencies between the stereo views, focus on relevant information, and leverage contextual relationships within the visual data.

Section 5: Performance evaluation of H-Net

5.1 Network convergence

We trained the H-Net for 800 epochs, with each epoch consisting of 80 iterations, resulting in a total of 64,000 iterations. We have carefully analyzed the training process and plotted the training loss curve based on the iterations, as shown in Figure S8. The graph clearly shows the trend of the training loss decreasing over time, indicating the convergence of our model during the training process. Starting from an initial training loss of 113, we observed a significant reduction in the loss as the training progressed. The training loss steadily decreased and eventually converged to around 0.3.

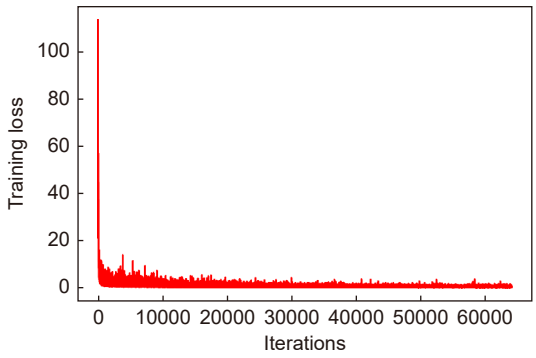

Fig. S8 | The training loss curve of H-Net based on iterations.

5.2 Evaluation metrics

We use the percentage of the three-pixel-error, the percentage of the one-pixel-error, the end-point error, and runtime to evaluate the network performance. The percentage of the three-pixel-error displays the percentage of predicted disparity pixels whose absolute difference from the ground-truth disparity value is greater than 3. The absolute difference map $\mathcal{D}_{diff}(D, \widehat{D}) = \left\{disp_{diff_n}\right\}_{n=1}^{N_{\text{total}}}$ is specifically calculated by Eq. (S16).

$$disp_{diff_n} = \left| d_n - \widehat{d}_n \right| \,, \tag{S16}$$

The percentage of three-pixel-error is further calculated as shown in Eq. (S17).

$$ThreePixelErr\left(D, \widehat{D}\right) = \frac{N_{disp_{diff}>3}}{N_{\text{total}}} \times 100\% \,, \tag{S17}$$

where D is the ground truth disparity map, $\widehat{D}$ is the predicted disparity map, $disp_{diff_n}$ is the absolute difference between ground truth and predicted disparity value for pixel n, N_{total} is the total number of pixels in the disparity map D (and $\widehat{D}$, and $\mathcal{D}_{diff}$), d_n is the ground truth disparity data for pixel n, and $\widehat{d}_n$ is the predicted disparity data for pixel n, $N_{disp_{diff}>3}$ is the number of pixels whose $disp_{diff_n}$ is greater than 3. For one-pixel-error, the number of pixels to be counted $N_{disp_{diff}>1}$ is

the number of pixels whose $disp_{diff_n}$ is greater than 1.

End-point error is the mean absolute difference for all pixels between the estimated and ground-truth disparity maps. The specific calculation is demonstrated in Eq. (S18).

$$EndPointErr\left(D,\widehat{D}\right)=\frac{1}{N_{\text{total}}}\sum_{n}^{N_{\text{total}}}\left(d_n-\widehat{d}_n\right)\ , \tag{S18}$$

5.3 Performance improvement evaluation

To quantify the improvements of our H-Net, we compare it with the conventional block matching algorithm and two advanced neural network methods, PSMNet[S4] and Unimatch[S5], on the disparity computation accuracy on our homemade test set derived from our meta-lens system. In this comparison, PSMNet and Unimatch all use the open-source trained weights provided by their authors. Our H-Net and PSMNet were all trained on the KITTI 2012 dataset. Because the performance of Unimatch trained on KITTI is relatively poor, we additionally compared its performance based on the Middlebury dataset (its best performance).

1) Test set preparation

The test set on meta-lens contains 31 stereo image pairs with 31 ground-truth disparity maps. The specific experimental setup of the test set collection is demonstrated in Fig. S9(a). A textured pattern (as shown in Fig. S9(b)) was attached to the surface of a flat board. The flat board moved from a distance of 150 mm to 450 mm in 10 mm steps. In the range of 150 to 450 mm, objects can be clearly imaged, minimizing the adverse effects of imaging quality problems such as defocusing on the test. The distance refers to the separation length between the binocular meta-lens and the flat board. We captured images every time the flatboard moved. For each image, all the disparity values in its disparity map are the same because the imaging object is a uniform surface with the same depth. Therefore, we derive 31 stereo (left and right) image pairs with different depth-disparity pairs. The ground truth disparity map is derived from the depth calculation formula Eq. (S19).

$$depth=\frac{fb}{ps\cdot\left|\widehat{D}+U_{\text{offs}}+O_{\text{offs}}\right|}\ . \tag{S19}$$

In the depth calculation Eq. (S19), U_{offs} in our system is 0, $O_{\text{offs}}<0$ and $\widehat{D}<|O_{\text{offs}}|$. Therefore, Eq. (S19) could be simplified as Eq. (S20).

$$depth=-\frac{fb}{ps*\left(\widehat{D}+O_{\text{offs}}\right)}\ , \tag{S20}$$

Therefore, $\widehat{D}$ could be expressed as Eq. (S21).

$$\widehat{D}=-\frac{fb}{ps*depth}-O_{\text{offs}}\ , \tag{S21}$$

Through Eq. (S21), we could obtain the computational ground truth disparity data for each depth in the range of 150 to 450 mm, as displayed in Fig. S9(c). The computational disparity data were further validated by manual calibration. For each image in the test set, the corresponding feature point pixels are found manually, and their corresponding pixel displacements are consistent with the calculated ground truth disparity data.

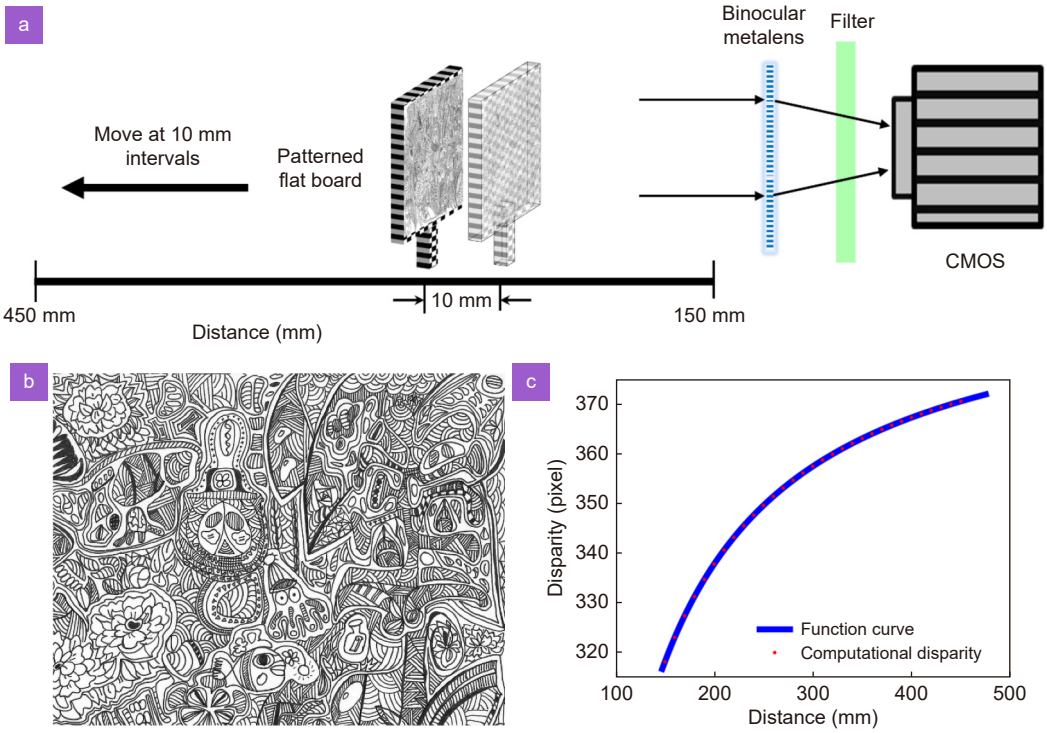

Fig. S9 | Configuration of the test set captured by meta-lens system. (**a**) The experimental setup for the image derivation of the test set. A patterned flat board moves from a distance of 150 mm to 450 mm in 10 mm steps. The definition of distance is the length from the plane of the flat board to the binocular meta-lens. (**b**) The pattern on the flat board, which is rich in texture. (**c**) The relationship between ground truth disparity data and distance according to the Eq. (S21). The blue line represents the function curve. The red dots are the ground truth disparity data corresponding to the images taken at distances ranging from 150 mm to 450 mm.

2) Comparison analysis

As presented in Table S1, our H-Net demonstrates superior performance compared to other methods across three evaluation metrics, including 1PE, 3PE, and EPE, over the entire test dataset. Generally, the 3PE metric is widely employed to assess the effectiveness of stereo-matching algorithms. We additionally employ the 1PE metric to further evaluate the algorithm's accuracy and robustness. Our method achieves an outstanding 1PE of 18.839%, surpassing that of other algorithms. This outcome substantiates the significant accuracy improvements brought about by the incorporation of the H-Module in the calculation of disparities.

Table S1 | Evaluation of different methods on the test set derived from our meta-lens system. We use the percentage of the three-pixel-error (3PE), the percentage of the one-pixel-error (1PE), the end-point error (EPE), and runtime for total test set evaluation. The results for the objects at 250 mm, 350 mm, and 450 mm are specifically listed for item comparison. All the results are tested on the Nvidia GeForce RTX 3090 GPU.

Method	Test Set on Meta-Lens									Runtime (s)
	250 mm		350 mm		450 mm		Total			
	3PE (%)	EPE	3PE (%)	EPE	3PE (%)	EPE	1PE (%)	3PE (%)	EPE	
Conventional Block Matching	0.088	0.690	**0.040**	0.741	**0.0**	0.805	36.886	0.181	0.877	~200
PSMNet	**0.0**	0.713	0.798	1.135	3.215	2.024	53.564	2.128	1.176	**0.144**
Unimatch (Middlebury)	0.126	1.269	0.201	1.133	1.392	1.579	75.904	2.902	1.604	0.503
Unimatch (KITTI)	61.951	6.769	51.384	4.894	60.566	6.474	84.622	58.694	6.455	0.503
Ours (H-Net)	**0.0**	**0.630**	0.170	**0.734**	**0.0**	**0.521**	**18.839**	**0.062**	**0.620**	0.147

Regarding runtime, H-Net exhibits comparable performance to the fastest PSMNet, with a mere 0.003 s difference in execution time. Considering that the introduction of the H-Module introduces additional parameters, it is reasonable for our algorithm to exhibit slightly slower performance. In contrast, the conventional method exhibits the longest runtime due to the trial-and-error hyperparameter selection process.

When comparing results for objects captured at distances of 250 mm, 350 mm, and 450 mm, our methods consistently outperform other approaches, except for a slightly inferior 3PE at 350 mm compared to the conventional block matching algorithm. However, the smaller EPE at 350 mm provides evidence of the enhanced robustness of our method compared to the conventional algorithm.

Figure S10 illustrates a comparative analysis of the disparity map computation results for objects located at distances of 250 mm, 350 mm, and 450 mm within the test set. Specifically, Fig. S10(a) showcases the original left image captured by our meta-lens. Figure S10(b-f) present the corresponding disparity maps obtained from the conventional block matching algorithm, PSMNet, Unimatch trained on the Middlebury dataset, Unimatch trained on the KITTI dataset, and our H-Net. Figure S10(g) represents the ground truth. Certain irregularities can be observed in the 250mm and 350 mm results generated by the conventional algorithm, as shown in Fig. S10(a). With the exception of Unimatch trained on the KITTI dataset, as depicted in Fig. 10(e), the outcomes from the other methods closely align with the ground truth. Our method provides better results with more uniform disparity distribution, especially in the 450 mm item.

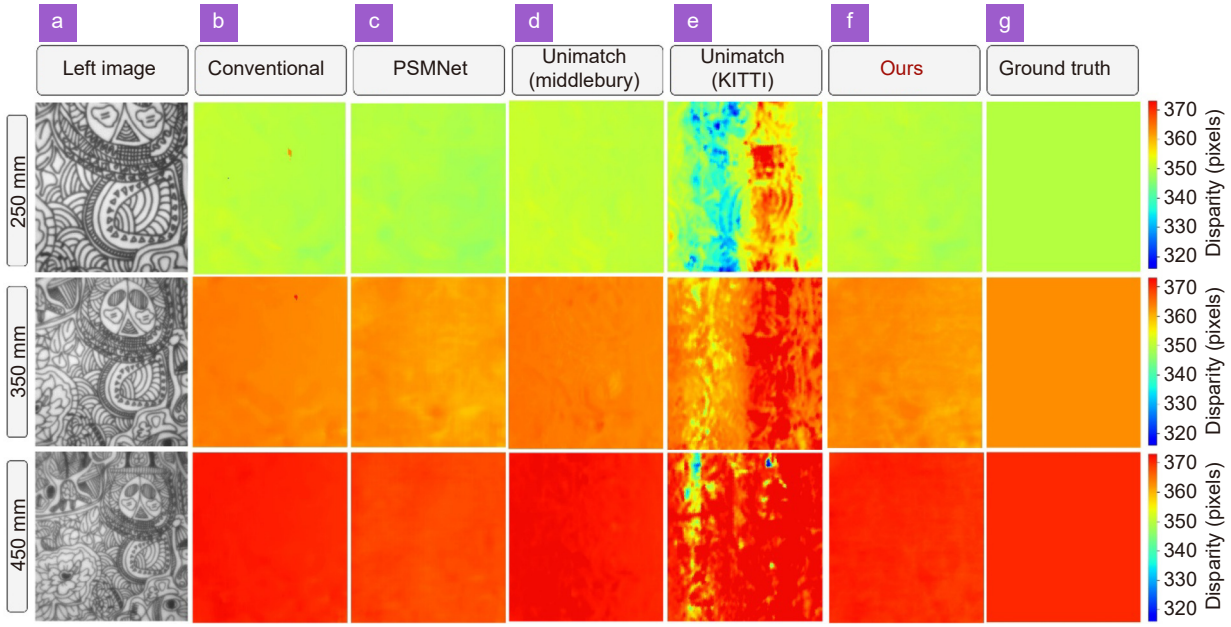

Fig. S10 | Disparity map computation result comparison on (**a**) the 250 mm, 350 mm, and 450 mm items in the test set among (**b**) Conventional, (**c**) PSMNet, (**d**) Unimatch trained on Middlebury dataset, (**e**) Unimatch trained on KITTI dataset, (**f**) Ours methods, and (**g**) Ground truth. The images in (a) are the corresponding left images captured by meta-lens.

5.4 Ablation study

We conducted the ablation experiments with and without H-Modules to evaluate H-Net. The default backbone of PSMNet[S4] was the basic architecture. We trained the H-Net and baseline on the stereo dataset KITTI 2012, which contains 194 training stereo image pairs with sparse ground-truth disparities obtained using LiDAR and 195 testing image pairs without ground-truth disparities. We further divided the whole training data into a training set (160 image pairs) and a validation set (34 image pairs). As our binocular meta-lens works under a single wavelength, the captured image is monochromatic. Therefore, the grayscale images of KITTI 2012 were adopted in model training. We use the percentage of the three-pixel-error and end-point error to evaluate the network performance.

As listed in Table S2, H-Net outperformed the baseline in both two quantitative indicators. In the baseline model (without the introduced ablation module), the Three Pixel Error is reported as 2.324%, and the End Point Error is 0.150. These metrics reflect the performance of the baseline model on the KITTI 2012 dataset. After introducing the

H-Module, the Three Pixel Error decreases to 1.258%, and the End Point Error decreases to 0.109. This reduction indicates that the incorporation of the H-Module has a positive impact on the model's performance, resulting in improved accuracy of the disparity map.

Table S2 | Evaluation of network with different settings. We calculated the percentage of the Three Pixel Error and End Point Error on the KITTI 2012 validation set.

Network setting		KITTI 2012	
Baseline	H-Module	Three Pixel Error (%)	End Point Error
√		2.324	0.150
√	√	**1.258**	**0.109**

The ablation experiment involving the H-Module demonstrates a significant improvement in the performance of the disparity estimation task on the KITTI 2012 dataset. The decrease in "Three Pixel Error" and "End Point Error" signifies enhanced accuracy and precision of the disparity map. These results validate the effectiveness of the H-Module and provide proof that the H-Module can capture contextual dependencies and enhance the understanding or analysis of the image.

References

S1. Albawi S, Mohammed TA, Al-Zawi S. Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)* 1–6 (IEEE, 2017); http://doi.org/10.1109/ICEngTechnol.2017.8308186.

S2. Luo WJ, Li YJ, Urtasun R, Zemel R. Understanding the effective receptive field in deep convolutional neural networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* 4905–4913 (ACM, 2016); http://doi.org/10.5555/3157382.3157645.

S3. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L et al. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* 6000–6010 (ACM, 2017); http://doi.org/10.5555/3295222.3295349.

S4. Chang JR, Chen YS. Pyramid stereo matching network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 5410–5418 (IEEE, 2018); http://doi.org/10.1109/CVPR.2018.00567.

S5. Xu HF, Zhang J, Cai JF, Rezatofighi H, Yu F et al. Unifying flow, stereo and depth estimation. *IEEE Trans Pattern Anal Mach Intell* **45**, 13941–13958 (2023).