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

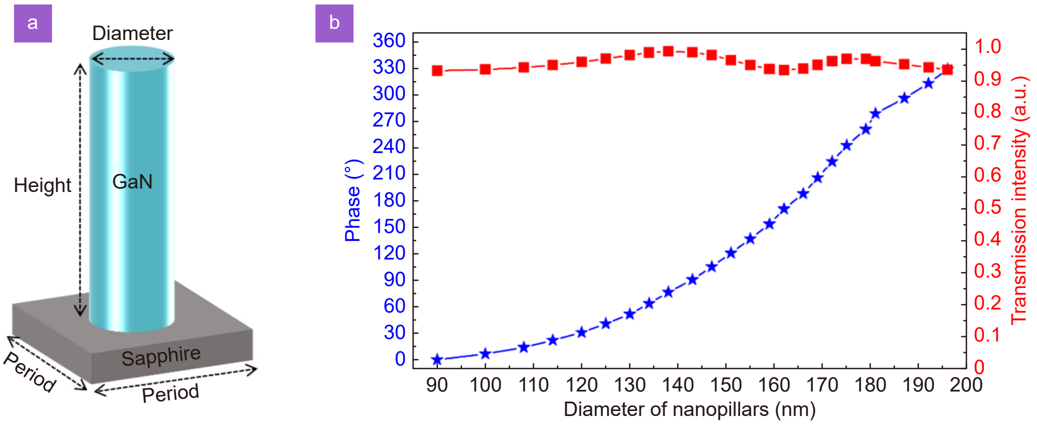

Fig. S1 | Meta-atom design and simulation of the binocular meta-lens. (**a**) The Schematic diagram of the GaN meta-lens fabrication process. (**b**) The phase modulation and transmission intensity of the meta-atom with various nanopillar diameters.

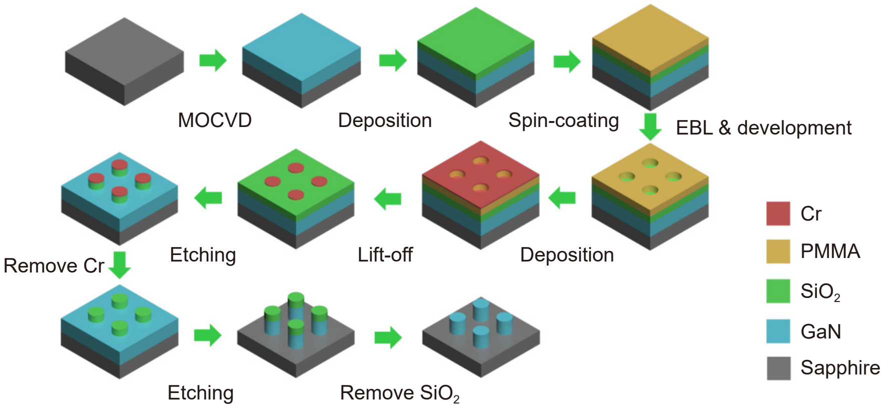

Fig. S2 | The Schematic diagram of the GaN meta-lens fabrication process.

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

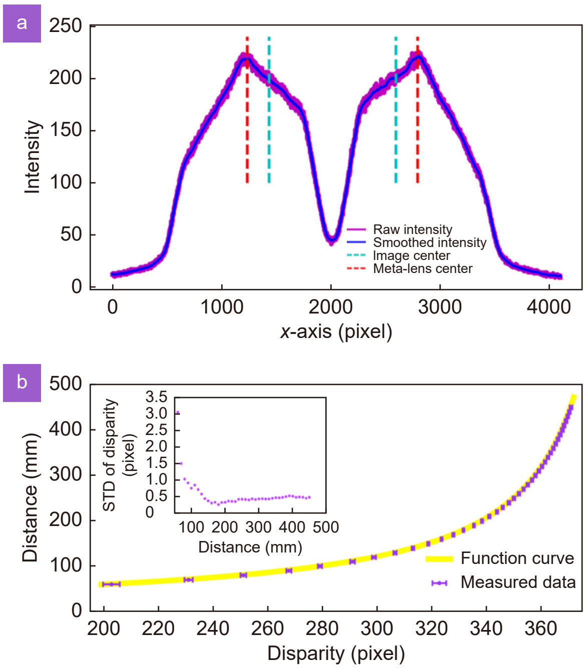

Fig. S3 | Depth calculation analysis based on our binocular meta-lens. (**a**) Intensity distribution along the cut line that crosses the two image centers when photographing a large white object. The distance between the red and cyan dashed lines is 0.5 O_{offs}. (**b**) The function relationship between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

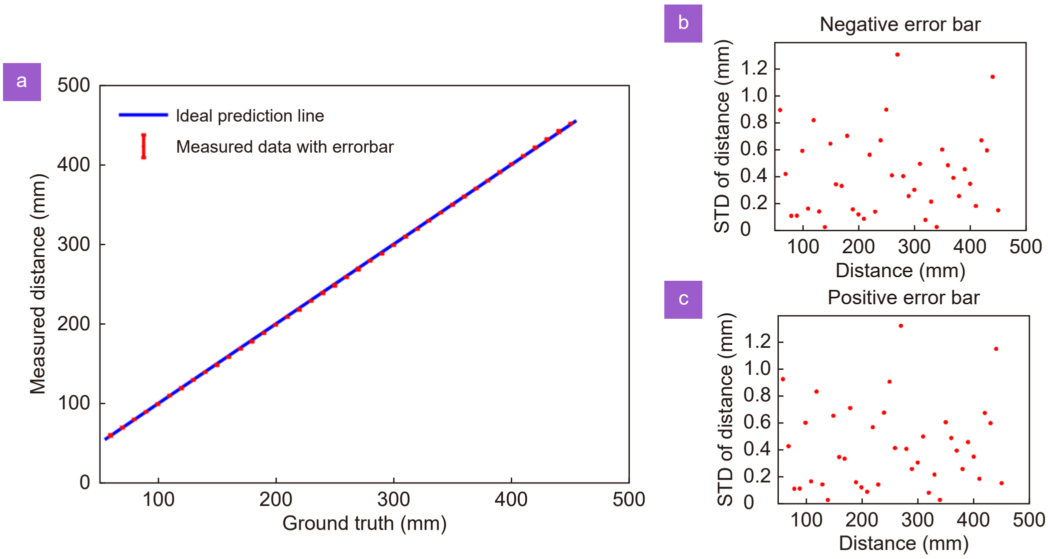

Fig. S4 | Depth resolution in the working range of 60 to 450 mm. (**a**) The measured distance with errorbar versus the ground truth distance. The ideal prediction line, depicted in blue, represents perfect agreement between measurements and ground truth values. The red dots represent the mean values 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,

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

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).

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

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

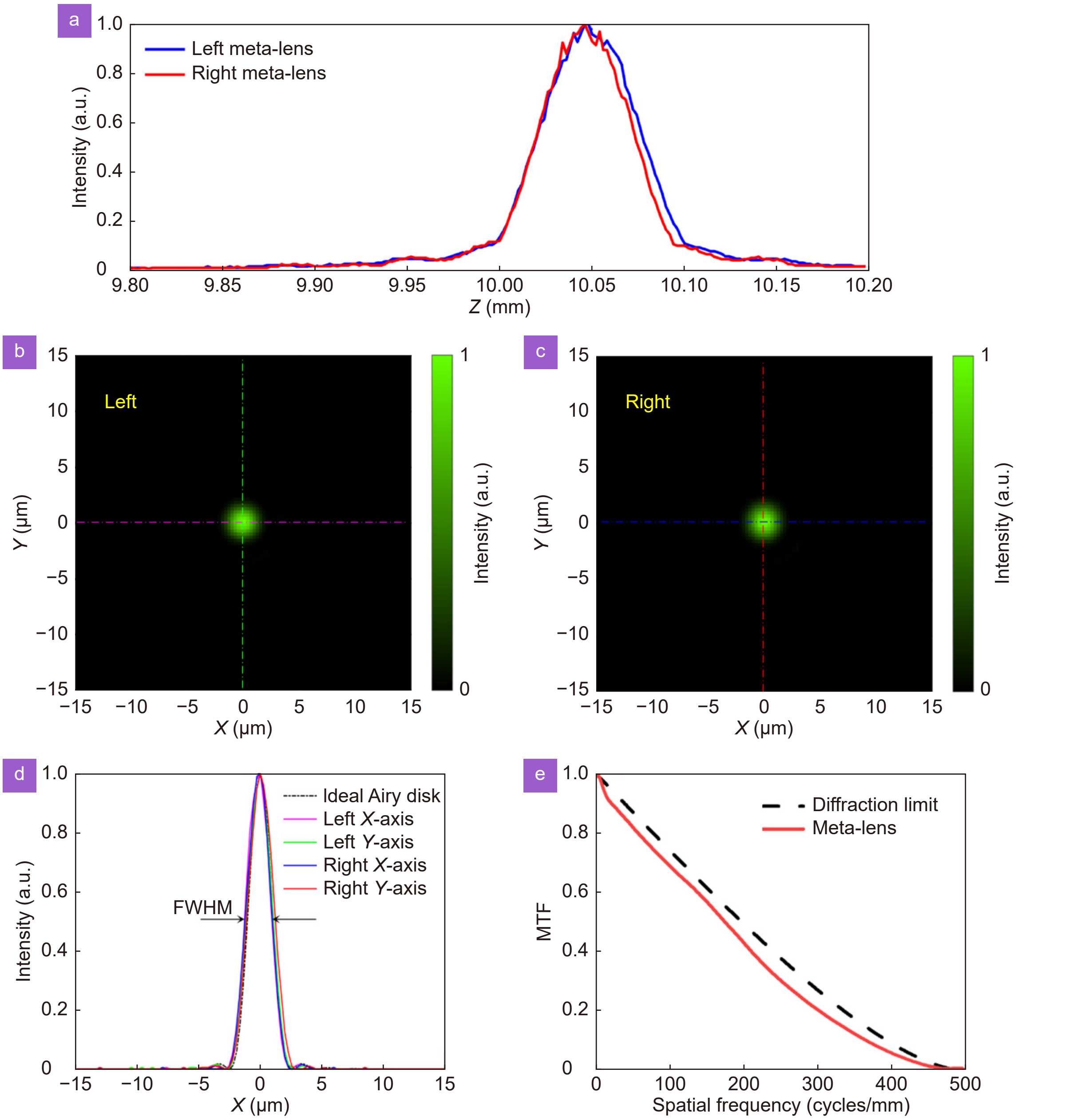

Fig. S5 | The optical performance of GaN meta-lens under 532 nm laser illumination. (**a**) The measured intensity profiles of left and right meta-lens along the optical axes (*z*-axis). (**b**) The measured focal spot image of the left meta-lens at the focal plane (*z* = 10.048 mm). (**c**) The measured 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.

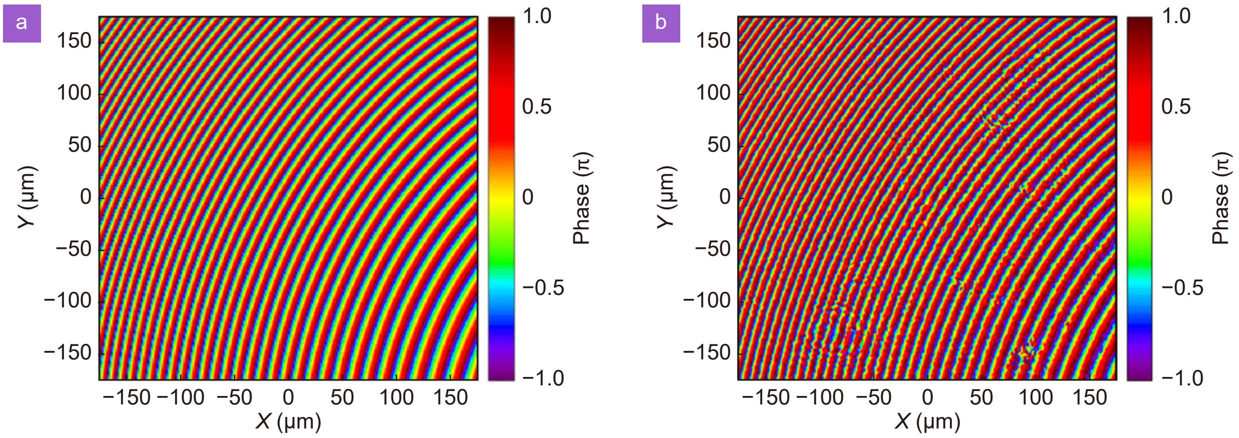

Fig. S6 | The phase characterization of GaN meta-lens. The calculated (**a**) and measured (**b**) phase profile at the edge region of the meta-lens.

Section 3: Configuration of the binocular meta-lens camera

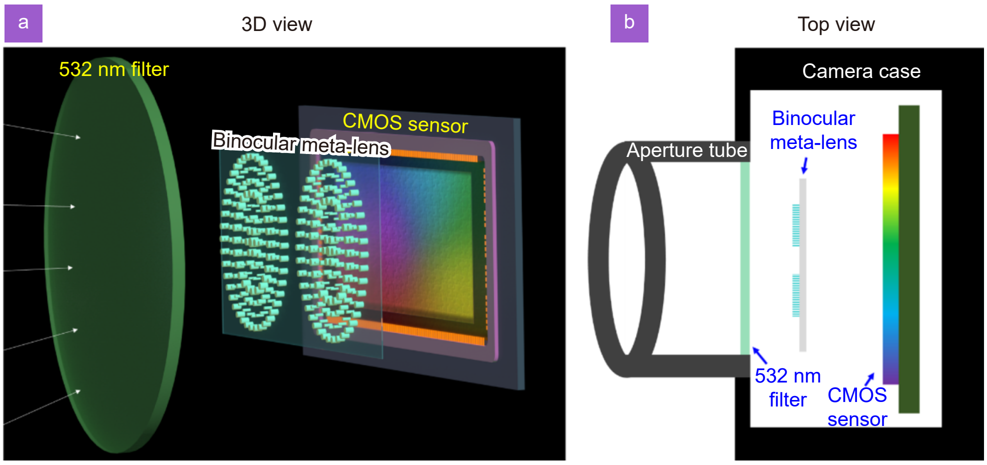

Fig. S7 | The configuration of the binocular meta-lens camera. The 3D view (**a**) and top view (**b**) of the binocular meta-lens camera.

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

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

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

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 $\mathcal{P}_l=\left\{p_{l_i}\right\}_{i=1}^{M\times N}$, and the right view, denoted as $\mathcal{P}_r=\left\{p_{r_i}\right\}_{i=1}^{M\times N}$. 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.

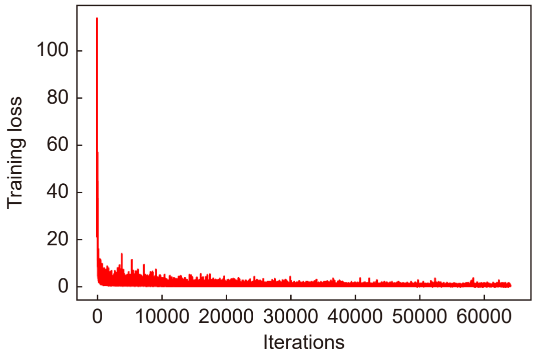

Fig. S8 | The training loss curve of H-Net based on iterations.

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.

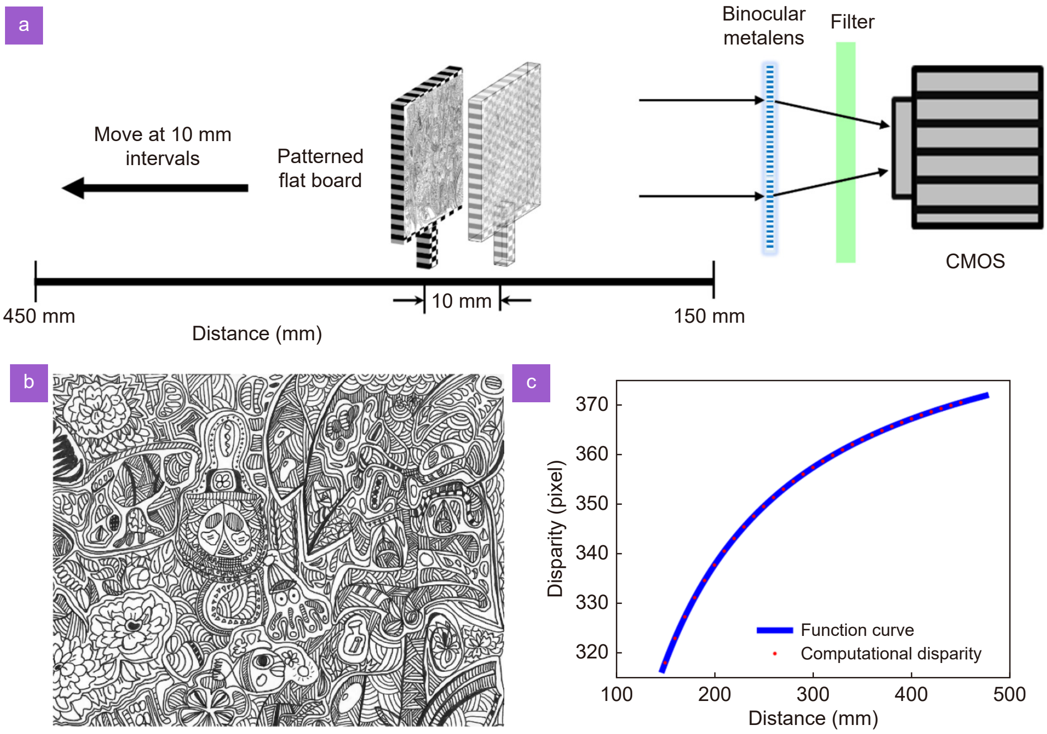

Fig. S9 | Configuration of the test set captured by meta-lens system. (**a**) The experimental setup for the image derivation of the test set. A patterned flat board moves from a distance of 150 mm to 450 mm in 10 mm steps. The definition of distance is the length from the plane of the flat board to 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.

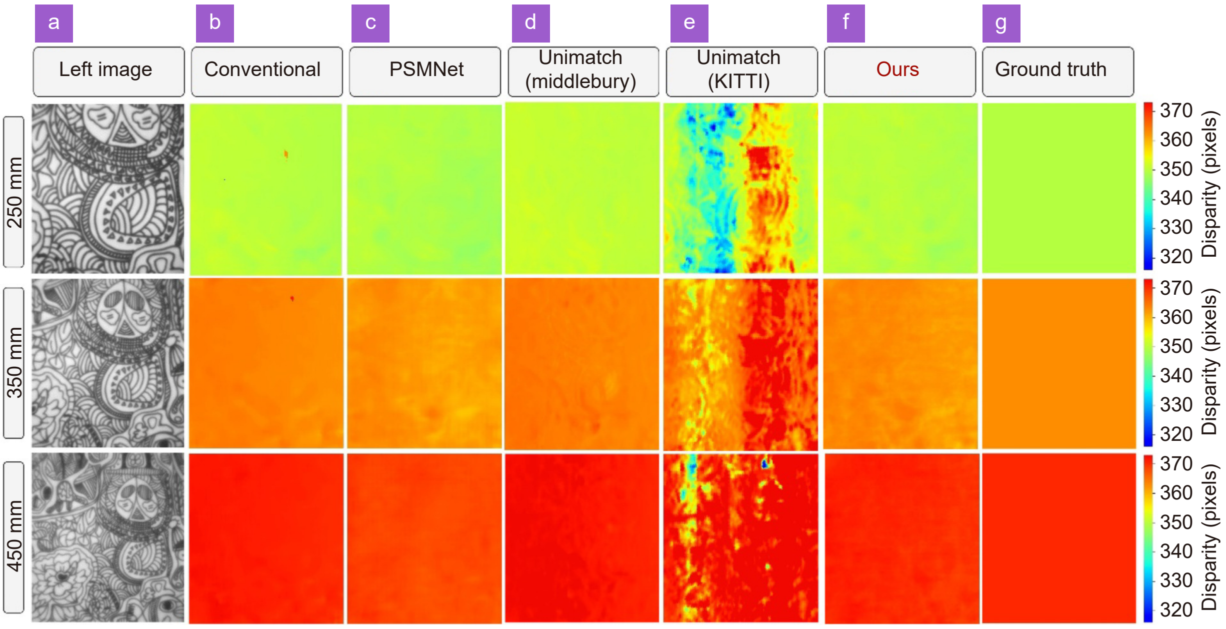

Fig. S10 | Disparity map computation result comparison on (**a**) the 250 mm, 350 mm, and 450 mm items in the test set among (**b**) Conventional, (**c**) PSMNet, (**d**) Unimatch trained on Middlebury dataset, (**e**) Unimatch trained on KITTI dataset, (**f**) Ours methods, and (**g**) Ground truth.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