https://doi.org/10.1038/s44310-025-00070-9

High-precision three-dimensional imaging based on binocular meta-lens and optical clue fusion

Yuzhou Song[1,2,8], Yifei Zhang[1,2,8], Xiaoyuan Liu[3,4], Takuo Tanaka[5,6,7], Mu Ku Chen[3,4] ✉ & Zihan Geng[1,2] ✉

Three-dimensional (3D) imaging plays a crucial role in autonomous driving, medical diagnostics, and industrial inspection by providing comprehensive spatial information. Metalens-based 3D imaging is highly valued for imaging applications thanks to its compactness, with enhanced precision remaining a key research pursuit. Here, we present an integrated high-accuracy 3D imaging system combining binocular meta-lens with an optical clue fusion network. Our innovation lies in the synergistic fusion of physics-derived absolute stereo depth measurements and machine learning-estimated relative depth through adaptive confidence mapping - the latter effectively addressing the inherent limitations of absolute depth estimation in scenarios with insufficient matching features. This hybrid approach achieves unprecedented precision of depth estimation (error <1%) while maintaining robust performance across feature-deficient surfaces. The methodology significantly expands viable detection areas and enhances measurement reliability, accelerating practical implementations of metalens-enabled 3D imaging.

Modern vision systems demand compact, high-performance solutions for object detection and depth perception—critical components in fields such as autonomous driving[1], robotics[2], and augmented reality[3]. Traditional stereo imaging setups using conventional lenses or multiple cameras tend to be bulky and require complex calibration procedures, limiting their integration into miniaturized platforms[4]. This has spurred the search for innovative optical designs and integration strategies that reduce system complexity while maintaining high-performance imaging and depth sensing. Recent advances in nanophotonics have introduced optical metasurfaces[5,6]—ultrathin, planar optical devices engineered with special designs—to overcome these challenges. By tailoring the phase, amplitude, and polarization of light at subwavelength scales, metasurfaces offer high design flexibility, enabling the creation of various optical elements such as lenses[7–10], sensors[11–13], holograms[14,15], resonators[16,17], and others. Moreover, they have been successfully utilized in diverse applications such as imaging[18–20], edge detection[21], encryption[22], wavefront engineering[23], and optical computing[24].

There has been extensive research on employing metasurfaces for depth sensing. The mainstream approaches in this area can be broadly divided into monocular[25–28] and binocular[12,29] systems. Monocular schemes typically employ passive imaging techniques such as dual-focal designs[25,26] and micro-lens arrays[30]. Although these methods eliminate the need for the calibration process inherent in binocular systems, they often face challenges such as imaging quality and inaccuracies of depth estimation[25]. In contrast, binocular systems demonstrate superior imaging quality while maintaining a compact design. The acquisition of depth information by binocular imaging depends on the perspective difference and pixel offset of the scene's texture information on the imaging surface. This depth information processing method will not work effectively for low-texture scenes. Recent years have witnessed a growing prevalence of neural networks in metasurfaces, not only for design optimization[31] but also for algorithmic implementations[32,33], resulting in significant enhancements to metasurface performance and capabilities. In particular, deep learning architectures—especially convolutional neural networks[34,35]—excel at extracting disparity information from stereo images, even in regions characterized by low texture or ambiguous features[36]. As demonstrated in our previous work, this synergy between meta-lens imaging and neural network processing enables precise, real-time depth estimation, thereby boosting object detection[37] and

[1]Institute of Data and Information, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, 518055, China. [2]Pengcheng Laboratory, Shenzhen, Guangdong, 518055, China. [3]Department of Electrical Engineering, City University of Hong Kong, Hong Kong SAR, 999077, China. [4]The State Key Laboratory of Terahertz and Millimeter Waves, City University of Hong Kong, Hong Kong SAR, 999077, China. [5]Innovative Photon Manipulation Research Team, RIKEN Center for Advanced Photonics, 351-0198 Saitama, Japan. [6]Metamaterial Laboratory, RIKEN Cluster for Pioneering Research, 351-0198 Saitama, Japan. [7]Institute of Post-LED Photonics, Tokushima University, 770-8506 Tokushima, Japan. [8]These authors contributed equally: Yuzhou Song, Yifei Zhang.
✉e-mail: mkchen@cityu.edu.hk; geng.zihan@sz.tsinghua.edu.cn

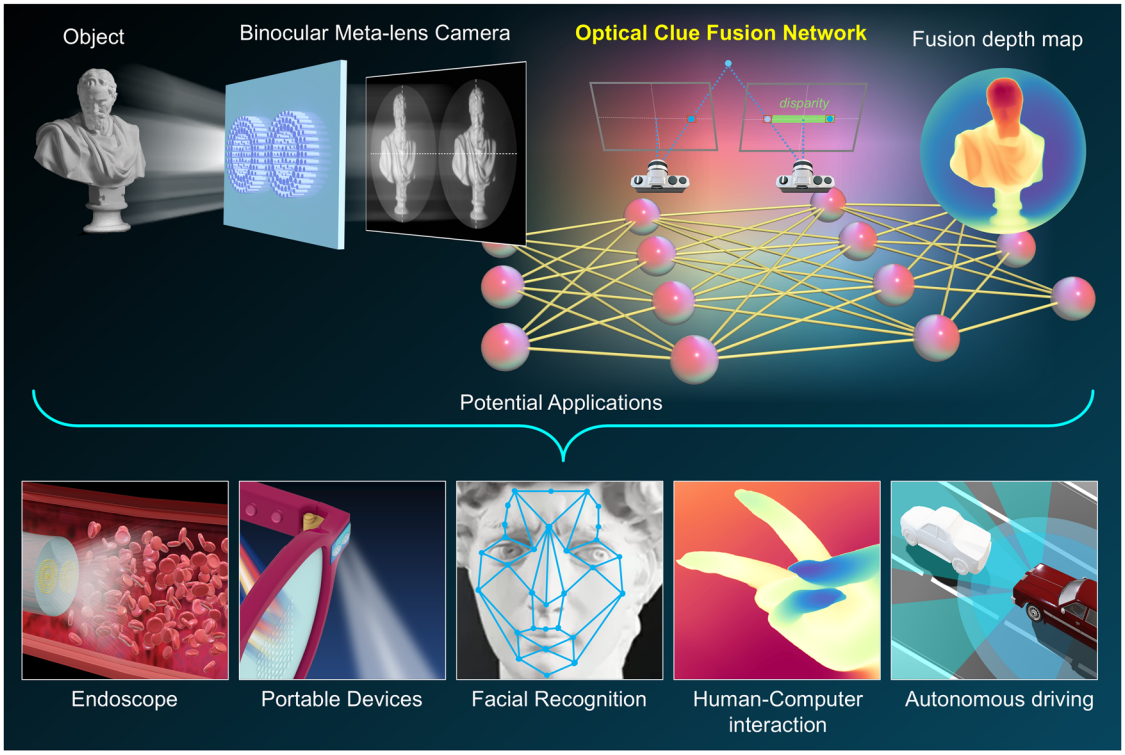

Fig. 1 | Schematic diagram of high-precision 3D imaging. The 3D imaging is based on binocular metalens and optical clue fusion, highlighting potential practical applications.

spatial understanding in complex environments[29]. In this study, we propose an Optical Clue Fusion Network (OCFN) that robustly integrates explicit stereo-geometric depth cues with implicit monocular depth priors for accurate depth estimation. Furthermore, we seamlessly incorporate this algorithm into a binocular metalens imaging system, as illustrated in Fig. 1. Rather than relying solely on stereo disparity, OCFN first refines the raw absolute depth map and then leverages a transformer-based monocular estimator to provide dense relative depth information. Our method yields a high-quality, dense absolute depth map through a two-stage, certainty-aware fusion process that includes calibrating the relative depth map via scaling and subsequently blending it with the cleaned stereo depth (weighted by Gaussian-smoothed confidence values). This sophisticated fusion strategy significantly enhances 3D scene perception, enabling more detailed and accurate depth representation. The compact and efficient nature of our binocular metalens system, combined with the powerful OCFN algorithm, opens new possibilities across numerous fields. Potential applications include immersive virtual reality experiences, responsive gaming environments, advanced 3D rendering, portable endoscopic devices, precise facial recognition systems, intuitive human-computer interaction interfaces, and reliable perception systems for autonomous driving.

Results

Binocular meta-lens

The binocular meta-lens consists of two identical meta-lenses, each with a diameter of 2.6 mm, separated by a horizontal distance of 4 mm. The phase profile for each individual meta-lens is designed as follows:

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

where λ is the operating wavelength (532 nm in our design), and f is the designed focal length (10 mm). We employ GaN cylindrical nanoantennas as meta-atoms. Each meta-atom provides distinct phase modulation, which depends on the feature size of the meta-atoms. The height of the meta-atom is fixed at 750 nm, while the diameter of cylindrical meta-atoms ranges from 90 nm to 196 nm, with a period of 260 nm, as shown in Fig. 2a. The polarization-independent phase modulation covers the entire 2π range by varying the diameter of the cylindrical GaN nanostructures. The phase shift and transmission intensity data are derived from numerical simulations using COMSOL Multiphysics®. The Simulation setup can be found in Section 1 of the Supporting Materials. Figure 2b presents the optical characteristics of the meta-atoms within the meta-device, obtained using COMSOL Multiphysics®. The transmittance fluctuates slightly across the entire range of diameter but generally remains around 0.95. As the diameter increases, the phase gradually increases from approximately 0 to 2π radians. The meta-atom array of the metalens is arranged according to the phase profile described in Eq. (1).

Figure 2c presents the photograph of the fabricated binocular meta-lens. The sample fabrication process is detailed in the Methods section. Figure 2d displays the scanning electron microscope (SEM) images of the meta-lens, confirming that no cracks or pores are present in the fabricated meta-atoms. The zoomed-in, tilted view of a meta-atom in Fig. 2e further reveals the well-collimated 750 nm-high structures achieved via precise etching. We constructed an optical setup to evaluate the performance of the binocular meta-lens. The experimental setup is described in Section 4 of the Supporting Materials and is illustrated in Fig. S3. The results of the scanning light field experiment are presented in Fig. 2f. The left eye meta-lens has a focal length of 9.962 mm, and the right eye meta-lens has 9.958 mm. At a wavelength of 532 nm, the measured focusing efficiency is 74%. The efficiency is defined as the ratio of the optical power within the focal spot area—where the central intensity falls to half its maximum value—to the total incident optical power across the meta-lens area. Additional parameters of the metalens are detailed in Section 3 of the Supporting Materials and illustrated in Fig. S2.

Furthermore, we integrated the binocular meta-lens onto a CMOS sensor, as depicted in Fig. 2g, and incorporated a 532 nm filter to achieve optimal performance. Although our proposed meta-lens exhibits a broadband response, its efficiency and the shape of its focal spot are

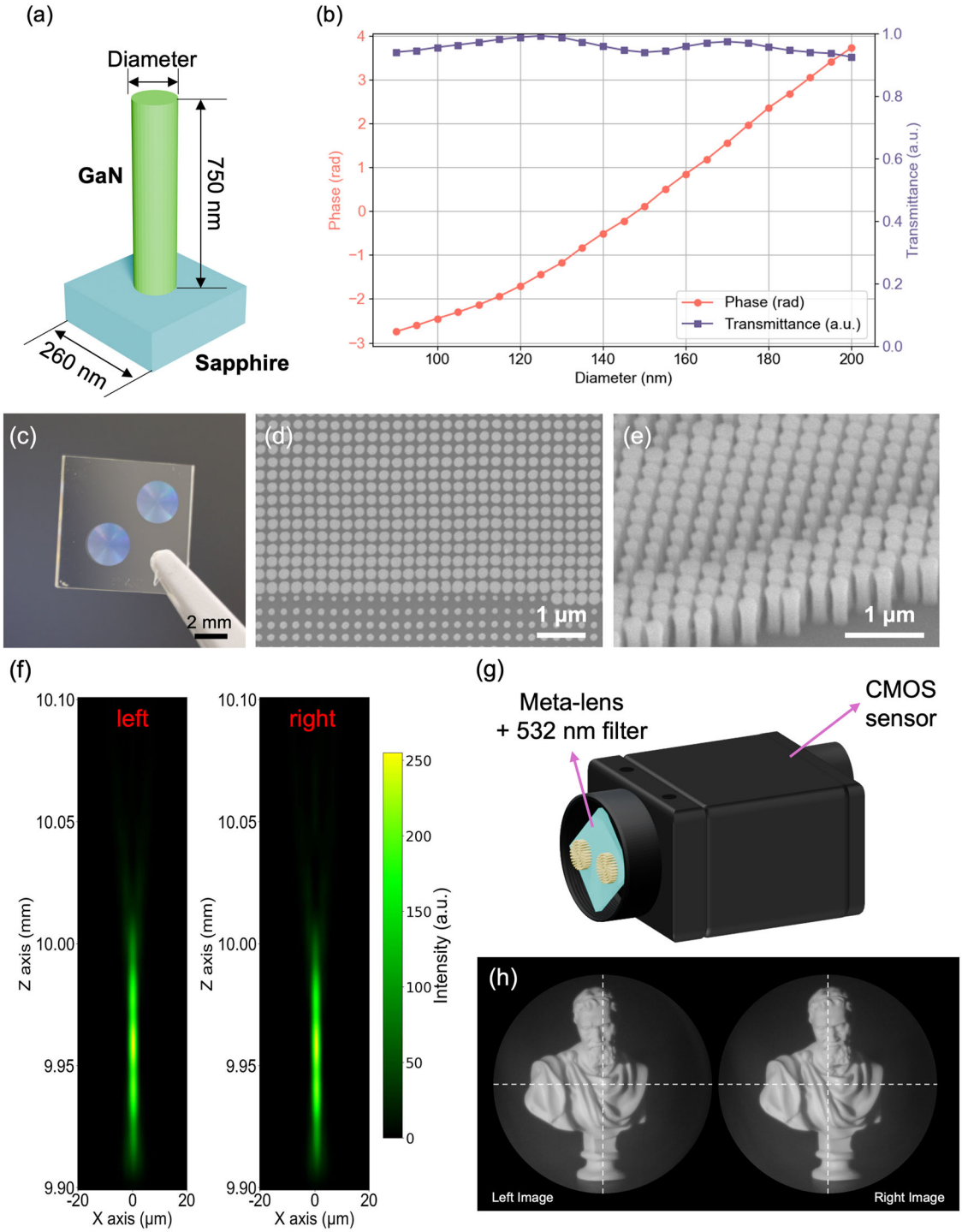

Fig. 2 | Schematic diagram of our binocular metalens system, showing its design, fabrication, and optical properties. (**a**) The schematic diagram of the meta-atom features a cylindrical GaN structure positioned on a sapphire substrate. (**b**) Optical characteristics of the meta-atom: the red curve shows the phase variation with respect to the diameter, while the purple curve shows the efficiency variation for the diameter. (**c**) Photograph of the fabricated binocular meta-lens. (**d**) SEM image of the meta-lens (scale bar: 1 μm). (**e**) Zoomed-in SEM image of the meta-lens (scale bar: 1 μm). (**f**) Optical scanning focusing profile results for the left and right eyes meta-lenses. (**g**) Schematic of the binocular meta-lens integrated with a CMOS sensor. (**h**) Raw single-shot image captured by the binocular system, simultaneously displaying left and right views.

affected when operating at wavelengths other than 532 nm. Detailed broadband experimental results can be found in our previous work[12,29,37]. Furthermore, the integration of broadband design into the system is feasible[38,39]. Figure 2h shows the raw image captured by our binocular imaging system, where the left and right images are acquired simultaneously in a single shot.

Optical Clue Fusion Network

The proposed Optical Clue Fusion Network (OCFN) is developed to provide high-quality and reliable 3D perception of the scene. As illustrated in Fig. 3, OCFN is designed to fuse two complementary sources of depth information: sparse, absolute depth obtained from stereo matching and dense, relative depth predicted by a pretrained monocular depth estimation model,

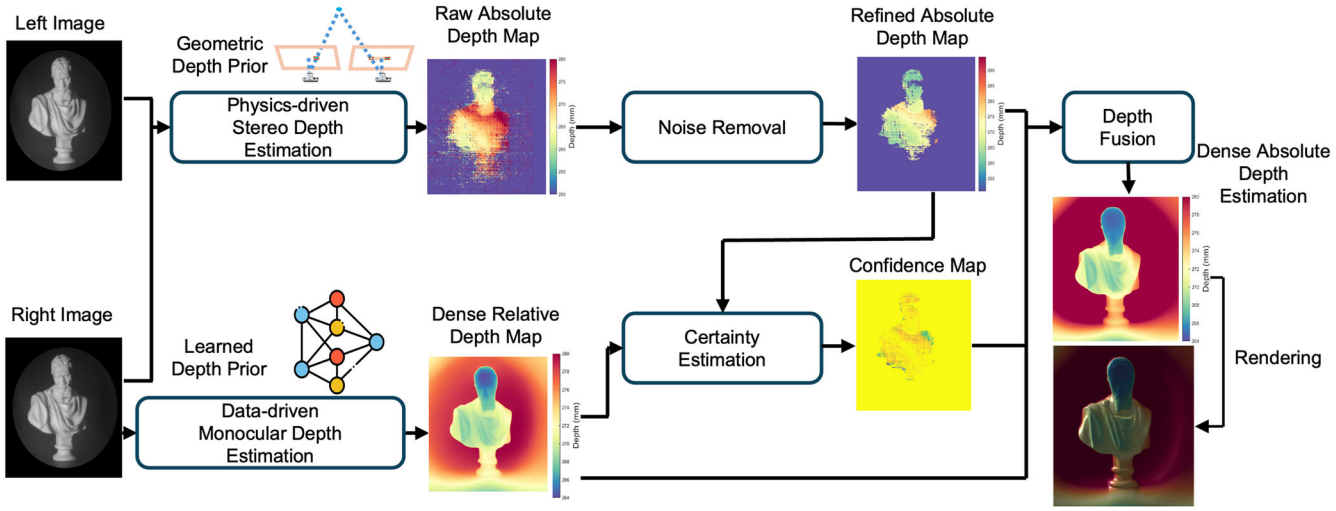

Fig. 3 | Details of the Optical Clue Fusion Network. A low-quality raw absolute depth map is initially computed from the stereo observations using the binocular imaging model. To enhance depth perception, a learning-based monocular depth estimator is applied to generate a high-quality relative depth map, capturing implicit depth cues from the scene. These two depth maps are then fused using a certainty-aware fusion strategy, resulting in a dense and accurate absolute depth map. The final output is rendered as high-quality 3D models for applications such as virtual reality, gaming, and other 3D rendering tasks.

specifically DepthAnything[40], more details of the monocular depth estimation model could be found in Section 6 of Supporting Materials. The motivation behind OCFN is to overcome the limitations of each individual method—stereo depth is geometrically accurate but often incomplete or noisy in textureless or occluded regions, while monocular depth is dense and visually consistent but lacks real-world scale and may contain local distortions.

OCFN addresses this by first aligning the monocular depth map to the scale of the stereo depth using a linear fitting model based on reliable stereo points. Then, it constructs a confidence map to assess the agreement between the two estimates at each pixel. A certainty-aware blending strategy is applied, where the final depth at each pixel is determined by a weighted combination of the scaled monocular and interpolated stereo depth, with the weights derived from the confidence map. This two-stage fusion framework ensures that depth values are both complete and geometrically faithful. The benefit of OCFN is that it enables robust and high-fidelity dense 3D reconstruction even in challenging imaging scenarios, such as those involving low-texture surfaces or complex geometries. The formula derivation of OCFN is presented in detail in the Methods section.

The 3D Perception Results

A series of experiments are conducted to verify the effectiveness of the proposed system. The algorithm is executed on an Intel Core i9-9900 CPU and an NVIDIA RTX-3090 GPU. The average processing time for predicting a depth image with 1200×1200 pixels is 0.6 seconds. The storage requirement during the inference of OCFN is 2.9GB, including the DepthAnything model, input tensors, and fusion buffers. The experimental results are shown in Fig. 4. Traditional imaging-model-based stereo matching methods struggle to produce detailed and reliable depth maps, especially in challenging scenarios. This limitation arises from their strong dependence on local texture cues to establish correspondences between stereo image pairs. In regions with uniform brightness, repetitive patterns, or poor texture—such as plain surfaces or smooth objects—the lack of distinct visual features prevents the algorithm from reliably identifying matching points. Additionally, external factors such as insufficient illumination, limited transmission efficiency, and minor lens manufacturing errors further degrade matching accuracy, resulting in sparse, noisy, or structurally fragmented depth maps. In contrast, our proposed OCFN method leverages a learned monocular depth prior to compensating for these weaknesses and significantly enhances depth perception quality. To illustrate the depth structure intuitively, we use MATLAB's built-in visualization tool (surf function) to rendering the estimated depth maps into surface plots with proper lighting and viewing angles, as is shown in the final column. In Fig. 4a, our OCFN depth estimation identifies the occlusion relationships between the poker cards and successfully maintains the boundary of each card. Stereo depth estimation, however, only recovers the depth of each pattern on the cards and fails to preserve the overall shape of the cards. In Fig. 4b, two plaster busts are placed side by side. The depth results of the stereo-matching method cannot distinguish the 3D details of the plaster busts. By our OCFN, detailed 3D information can be obtained. In Fig. 4c, a piece of paper printed with "THU" is attached to a tilted board and placed in front of the camera. Since most of the paper is plain white, there are insufficient corresponding points in the stereo estimation. As a result, only the edges of the letters can be inferred for depth information. In comparison, OCFN effectively fuses partial stereo-depth information with the learned monocular perspective prior, successfully revealing the incline of the paper. Even the slight depth difference between the paper and the board can be detected. Thanks to the additional learning-based depth prior in OCFN, the resulting predictions capture more details and more faithfully reflect the photometric properties of the original scene. As shown in the rendered 3D models, the OCFN predictions are both geometrically accurate and visually pleasing, which is crucial for downstream applications.

Measurement Accuracy and Resolution Analysis

Figure 5a visualizes the measurement accuracy analysis of our proposed system. We sequentially placed spade poker cards representing 8, 9, 10, J, Q, and K on a platform, with a 10 mm gap between each card. The first poker card is positioned at 220 mm, while the meta-lens is located at 0 mm. The depth reconstruction results demonstrate that the proposed OCFN method achieves high accuracy. The depth of each reconstructed card is smooth and uniform. Numerically, the predicted depth shows less than a 1% deviation from the ground truth depth of each card, which is acceptable, especially when considering potential errors in card positioning. This accuracy arises from the hybrid nature of our OCFN, which uses physics-based stereo depth estimation as the cornerstone of 3D perception. Other than solely relying on a learned-based depth estimation method, this hybrid framework allows for more interpretability and reliability of depth prediction.

Figure 5b presents the measurement resolution analysis of the proposed system. In this configuration, a chessboard pattern is affixed to a white panel, which is mounted on a motorized translation stage (Thorlabs PT1-Z9) with a minimum precision of 0.2 μm. The initial distance between the screen and the metalens is approximately 212.5 mm. Experiments are performed by translating the screen backward in 30 μm increments, starting

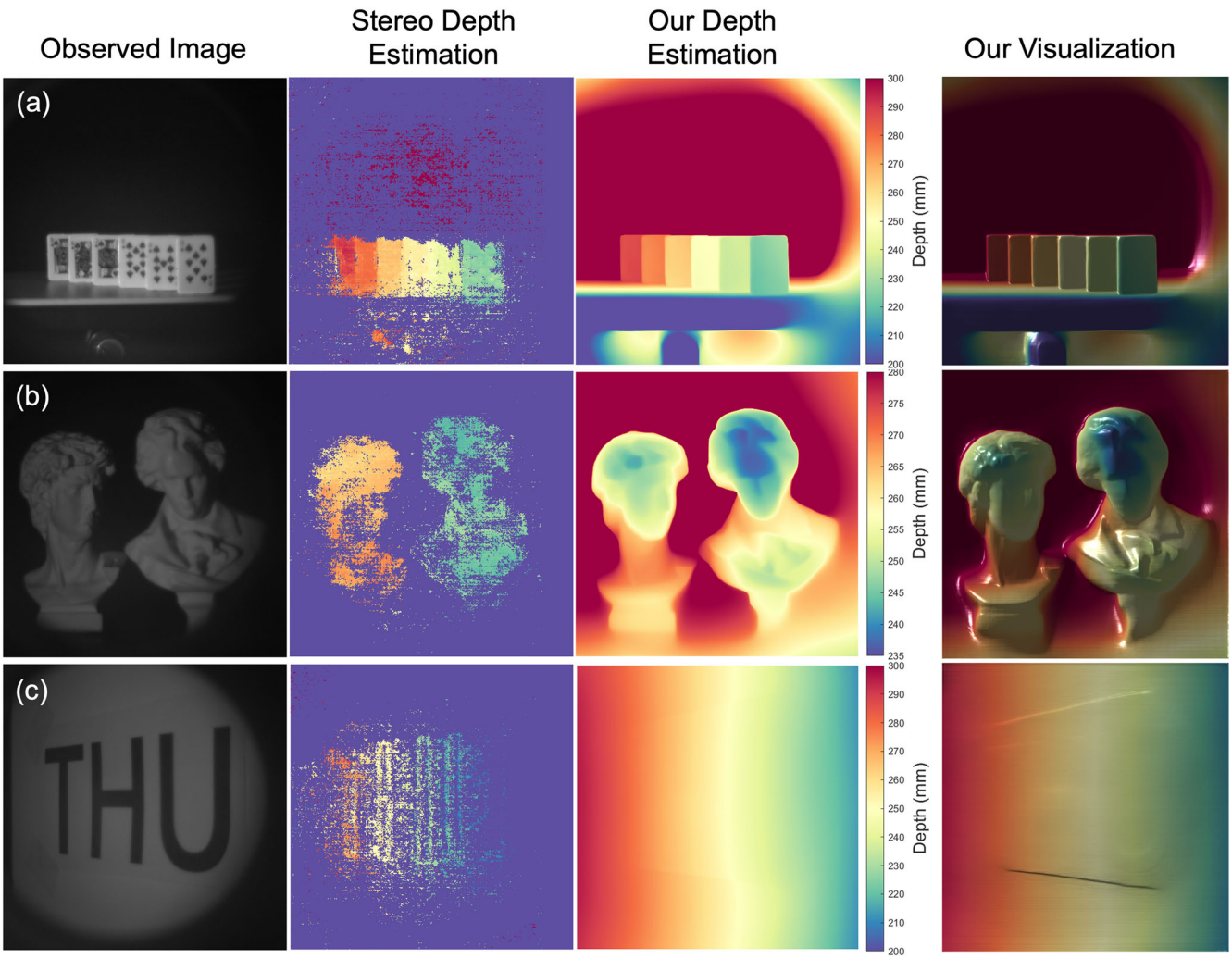

Fig. 4 | Real-life testing results for various scenarios. The leftmost column shows the raw captured right image. The second column presents the result from the physics-based stereo depth estimation method. The third column displays the outcome of the OCFN method. The rightmost column shows the rendering results derived from our OCFN depth estimations, which can be applied in gaming and virtual reality applications. (**a**) A scenario with several poker cards placed at different distances; (**b**) A scenario with two plaster busts; (**c**) A scenario with an inclined printed paper.

from 0 μm up to 990 μm, resulting in a total of 34 test points. The detailed experimental setup is described in Section 5 of the Supporting Materials and illustrated in Fig. S4. We use the proposed OCFN to keep track of the translation process and estimate the depth of the screen. The error bars indicate the standard variations of depths in different screen regions. The results show that the proposed system can successfully tell the difference between two planes with a 30 μm shift. Therefore, the proposed system provides a sensitive tool for precise manufacturing and other 3D perception applications with high resolution requirements. We also visualize a few select depth estimation results of the moving screen. The result is smooth with low fluctuations, showcasing the reliability and robustness of the proposed metalenses-based perception system and the corresponding OCFN depth estimation method.

Discussion

In this work, we integrate a binocular meta-lens with an Optical Clue Fusion Network (OCFN) to achieve high-quality 3D depth perception. The fabricated binocular meta-lens demonstrates a focal length deviation of less than 0.5% from the 10 mm design target and delivers a high focusing efficiency of 74% at 532 nm. Moreover, it can be seamlessly integrated with CMOS sensors to enable single-shot 3D imaging. The system substantially improves overall depth coverage by merging stereo-derived absolute depth with a transformer-based monocular relative depth, even in texture-less regions. The system employs a two-stage fusion process. First, it calibrates the monocular depth using linear scaling parameters. Subsequently, a 9×9 Gaussian-smoothed confidence map is applied to weight and fuse the monocular and stereo information, producing a dense depth map with high geometric accuracy and detail. The system can reliably capture fine boundary details and process 1200×1200 images in 0.6 seconds. The image reconstruction can be further accelerated with several strategies, including 1) using smaller monocular depth estimation network variants, 2) reducing input resolution, and 3) employing model quantization and runtime optimizations such as TensorRT. With these approaches, we believe real-time 3D sensing is feasible. Measurement evaluations show that the reconstructed depths deviate by less than 1% from ground truth, and the system can detect shifts as small as 30 μm, underscoring its potential in high-precision 3D measurement applications. For a practical demonstration of the system's capabilities, hand gesture 3D reconstruction examples are provided in Section 7 of the Supporting Materials, and the results are shown in Fig. S5. It demonstrated gesture recognition and human-computer interaction applications that may be used in portable devices in the future. This work develops a precise 3D imaging algorithm based on the binocular meta-lens camera. The high-precision 3D perception and rapid modeling will enable mobile phone face recognition, medical surgery navigation, autonomous driving obstacle avoidance, and industrial intelligent quality inspection, promoting intelligent upgrades in multiple fields.

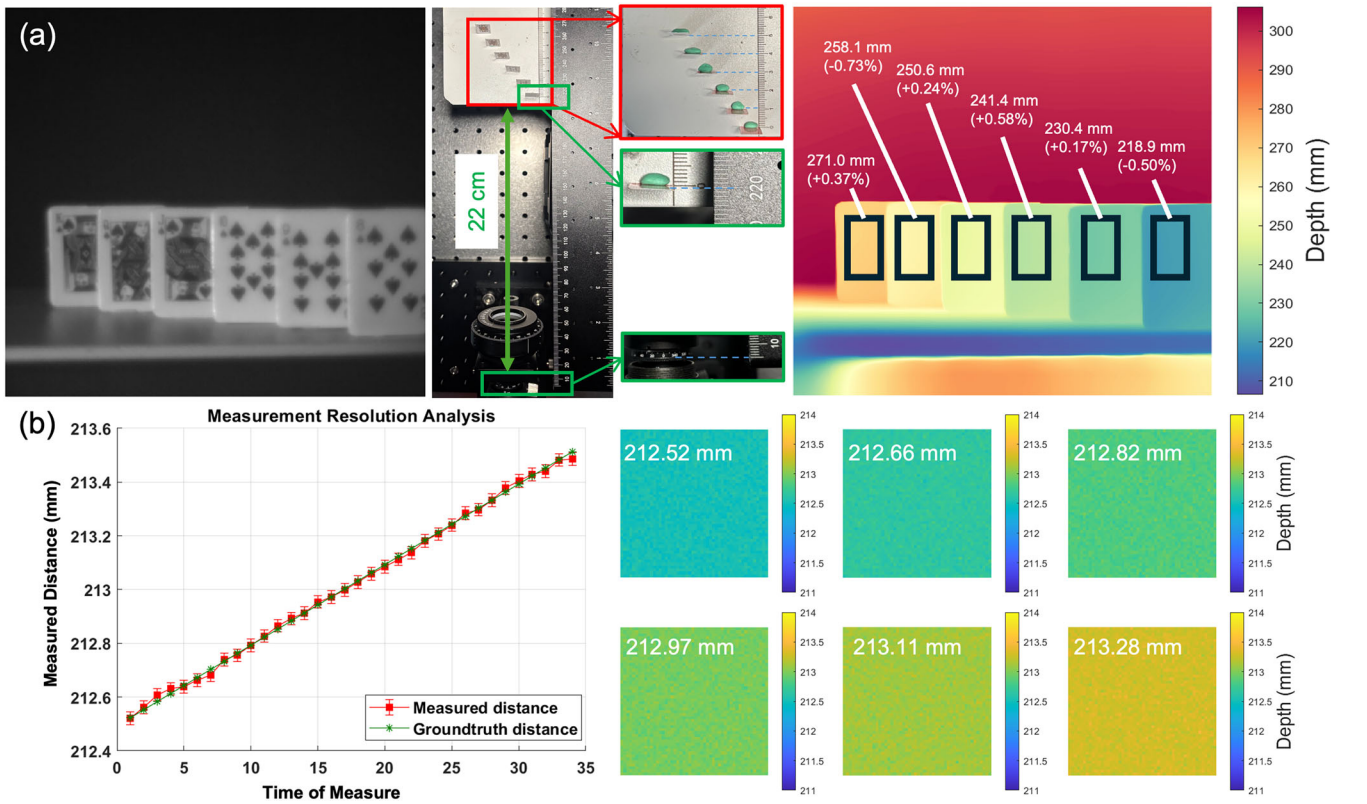

Fig. 5 | Measurement accuracy and resolution analysis. (**a**) We sequentially placed spade poker cards representing 8, 9, 10, J, Q, and K on a platform with a 10 mm gap between each card. The first poker card is positioned at 220 mm, while the meta-lens is placed at 0 mm. We show the measured depth of each card. (**b**) Depth resolution test of our system. Experiments are conducted by translating the screen backward in 30 μm increments, starting from 0 μm up to 990 μm, resulting in a total of 34 test points. The measured distances, along with a few selected recovered depth images of the screen, are visualized.

Methods

Sample fabrication

A 750-nanometer Gallium Nitride (GaN) layer is deposited on a sapphire substrate by Metal-Organic Chemical Vapor Deposition (MOCVD), followed by a 200-nanometer Silicon Dioxide (SiO_2) hard mask deposited using an Electron-gun evaporator. Polymethyl Methacrylate (PMMA) photoresist layer is spin-coated, pre-baked at 180 degrees Celsius, and patterned by Electron Beam Lithography (EBL). The sample is developed in Methyl Isobutyl Ketone/Isopropyl Alcohol (MIBK/IPA) and rinsed in Isopropyl Alcohol (IPA). A 40-nanometer Chromium (Cr) layer is deposited, and the pattern is transferred to the Cr layer via a lift-off process. The SiO_2 layer with the Cr metal mask is etched using Inductively Coupled Plasma - Reactive Ion Etching (ICP-RIE) with Carbon Tetrafluoride (CF_4). The patterned SiO_2 layer is the hard mask for the high aspect ratio GaN nanostructures etching. An ICP-RIE process with Argon/Chlorine (Ar/Cl_2) transfers the pattern to the GaN layer. Finally, the residual SiO_2 is etched away with the Buffered Oxide Etch (BOE), resulting in GaN nanostructures on the sapphire substrate. A schematic illustration of this nanofabrication process is provided in Fig. S1.

Details of the Optical Clue Fusion Network

OCFN integrates both explicit geometric priors and implicit learned priors to extract depth information. Using a certainty-aware fusion strategy, the depth maps derived from these dual priors are combined into a single dense absolute depth map, which serves as the final output. OCFN leverages the depth-disparity relation between binocular metalenses as an interpretable base for 3D depth perception. As the two metalenses observe the scene from slightly different viewpoints, corresponding points in the left and right images are offset by a certain amount, referred to as disparity. By measuring the disparity between these corresponding points, the depth of objects in the scene can be computed. This relation is given by:

$$Z = \frac{f\,B}{\delta} \tag{1}$$

where Z is the distance from the observer to the object, f is the focal length of the camera, B is the distance between the two metalens, and δ is disparity, which can be calculated with stereo matching algorithms.

OCFN employs a certainty-aware fusion strategy to fuse the depth information from two priors. The raw absolute map from the stereo depth is sparse and noisy but also contains absolute depth information. On the other hand, the dense relative depth map from learning-based monocular depth visually aligns well with the observed 2D image, but the depth information is measured on a relative scale and can be distorted. OCFN gradually fuses these two distinct depth maps in two stages. In the first stage, the raw absolute depth image obtained from stereo matching is refined by removing outlying data points, which are typically erroneous and unreliable. In our context, outlying points refer to small, isolated regions in the stereo depth map that are not spatially connected with other valid depth estimates. These often result from false stereo correspondences in textureless or occluded regions. To automatically detect and eliminate such noise, we apply a connected-component analysis to the binary depth mask using MATLAB's bwconncomp function. Regions with an area smaller than a predefined threshold (e.g., 200 pixels) are discarded, as they are considered statistically insignificant and likely to represent spurious matches. This filtering process is fully automated and removes the need for manual intervention, thus ensuring objectivity and consistency in the generation of the cleaned sparse absolute depth. The remaining valid points are then used as ground truth to fit a linear scaling model that maps the relative monocular depth to an absolute scale. This involves the estimation of a scaling factor a, and an

offset b. The process is given by:

$$a = \frac{\sum_{i=1}^{N}\left(\mathrm{D}_{\mathrm{relative}}(i) - \overline{D_{relative}}\right)\left(D_{abs}(i) - \overline{D_{abs}}\right)}{\sum_{i=1}^{N}\left(\mathrm{D}_{\mathrm{relative}}(i) - \overline{D_{relative}}\right)^2} \quad (3)$$

$$b = \overline{D_{abs}} - a\overline{D_{relative}} \quad (4)$$

where N is the number of sparse depth data points in the cleaned raw and sparse absolute depth map D_{abs} from stereo. $D_{abs}\,(i)$ is the depth at point i. $D_{relative}\,(i)$ is the relative depth at point i. After determining the scaling factor a and the offset b, we apply them to the entire relative depth map to produce the scaled absolute depth D_{scaled}.

$$D_{scaled} = aD_{relative} + b \quad (5)$$

In the second stage, OCFN blends the scaled absolute depth D_{scaled} based on its consistency with the cleaned raw absolute depth D_{abs}. Given a datapoint i, if $D_{abs}(i)$ aligns well with $D_{scaled}(i)$, we consider this point in D_{scaled} to be more reliable. A pixel-wise confidence map of depth estimation can be obtained as

$$C_{scaled}(i) = \frac{1}{1 + \left|D_{\mathrm{abs}}(i) - D_{scaled}(i)\right|} \quad (6)$$

For non-existing points in $D_{abs}(i)$, the corresponding $C_{scaled}(i)$ are set to ones. For each pixel in D_{scaled}, we calculate its weight based on the confidence.

$$w_{scaled}(i) = GaussianSmooth\left(C_{scaled}(i)\right) \quad (7)$$

where *GaussianSmooth* is an 9×9 Gaussian kernel spread the confidence from known depth values to neighboring pixels, increasing the reliability of the depth estimates in the regions surrounding sparse depth points. Finally, weighted blending is performed to combine the information from $D_{abs}(i)$ and $D_{scaled}(i)$. The formula is given by:

$$D_{fused}(i) = w_{scaled}(i)D_{scaled}(i) + \left(1 - w_{scaled}(i)\right)Interp\left(D_{abs}(i)\right) \quad (8)$$

where *Interp* is the bilinear interpolating function that helps to fill the gaps in the cleaned sparse depth map $D_{abs}(i)$. With this certainty-guided fusion strategy, the depth information from both priors is sufficiently utilized, leading to a high-quality dense absolute depth map as the final output.

Data availability

The data supporting this study's findings are available from the corresponding author upon reasonable request.

Received: 27 February 2025; Accepted: 10 May 2025;
Published online: 01 July 2025

References

1. Qian, R., Lai, X. & Li, X. 3D object detection for autonomous driving: a survey. *Pattern Recognit.* **130**, 108796 (2022).
2. Robinson, N., Tidd, B., Campbell, D., Kulić, D. & Corke, P. Robotic vision for human-robot interaction and collaboration: a survey and systematic review. *ACM Trans. Hum.-Robot Interact.* **12**, 1–66 (2023).
3. Zhan, T., Yin, K., Xiong, J., He, Z. & Wu, S.-T. Augmented reality and virtual reality displays: perspectives and challenges. *iScience* **23**, 101397 (2020).
4. Blake, R. & Wilson, H. Binocular vision. *Vis. Res.* **51**, 754–770 (2011).
5. Peng, Y. et al. Metalens in improving imaging quality: advancements, challenges, and prospects for future display. *Laser Photonics Rev.* **18**, 2300731 (2024).
6. Kossowski, N. et al. Metrology of metasurfaces: optical properties. *npj Nanophotonics* **2**, 5 (2025).
7. Liu, B. et al. Metalenses phase characterization by multi-distance phase retrieval. *Light.: Sci. Appl.* **13**, 182 (2024).
8. Liu, Y. et al. Linear electro-optic effect in 2D ferroelectric for electrically tunable metalens. *Adv. Mater.* **36**, 2401838 (2024).
9. Choi, M. et al. Roll-to-plate printable RGB achromatic metalens for wide-field-of-view holographic near-eye displays. *Nat. Mater.* **24**, 535–543 (2025).
10. Kuang, Y. et al. Palm vein imaging using a polarization-selective metalens with wide field-of-view and extended depth-of-field. *npj Nanophotonics* **1**, 24 (2024).
11. Chen, M. K. et al. A meta-device for intelligent depth perception. *Adv. Mater.* **35**, 2107465 (2023).
12. Liu, X. et al. Underwater binocular meta-lens. *ACS Photonics* **10**, 2382–2389 (2023).
13. Go, G.-H. et al. Meta Shack–Hartmann wavefront sensor with large sampling density and large angular field of view: phase imaging of complex objects. *Light.: Sci. Appl.* **13**, 187 (2024).
14. Zhou, H. et al. Multi-fold phase metasurface holography based on frequency and hybrid decoupling polarizations. *Adv. Optical. Mater.* **13**, 2402303 (2025).
15. Wen, X. et al. Quasicrystal metasurface for optical holography and diffraction. *Light.: Sci. Appl.* **13**, 246 (2024).
16. Rybin, M. V. & Kivshar, Y. Metaphotonics with subwavelength dielectric resonators. *npj Nanophotonics* **1**, 43 (2024).
17. Deng, Q.-M. et al. Advances on broadband and resonant chiral metasurfaces. *npj Nanophotonics* **1**, 20 (2024).
18. Cheng, J. et al. Tunable meta-device for large depth of field quantitative phase imaging. *Nanophotonics.* **14**, 1249–1256 (2025).
19. Song, Y. et al. Three-dimensional varifocal meta-device for augmented reality display. *PhotoniX* **6**, 6 (2025).
20. Xing, Z. et al. Monolithic spin-multiplexing metalens for dual-functional imaging. *Laser Photonics Rev.* 2401993, https://doi.org/10.1002/lpor.202401993 (2025).
21. Zhou, Y. et al. Meta-device for field-of-view tunability via adaptive optical spatial differentiation. *Adv. Sci.* **12**, 2412794 (2025).
22. Zhang, F. et al. Meta-optics empowered vector visual cryptography for high security and rapid decryption. *Nat. Commun.* **14**, 1946 (2023).
23. Huang, S.-H. et al. Microcavity-assisted multi-resonant metasurfaces enabling versatile wavefront engineering. *Nat. Commun.* **15**, 9658 (2024).
24. Xu, D. et al. All-optical object identification and three-dimensional reconstruction based on optical computing metasurface. *Opto-Electron. Adv.* **6**, 230120–10 (2023).
25. Guo, Q. et al. Compact single-shot metalens depth sensors inspired by eyes of jumping spiders. *Proc. Natl. Acad. Sci.* **116**, 22959–22965 (2019).
26. Shen, Z. et al. Monocular metasurface camera for passive single-shot 4D imaging. *Nat. Commun.* **14**, 1035 (2023).
27. Yang, F., Lin, H.-I., Chen, P., Hu, J. & Gu, T. Monocular depth sensing using metalens. *Nanophotonics* **12**, 2987–2996 (2023).
28. Luo, Y. et al. Monocular metasurface for structured light generation and 3D imaging with a large field-of-view. *ACS Appl. Mater. Interfaces* **16**, 39906–39916 (2024).
29. Liu, X. et al. Stereo vision meta-lens-assisted driving vision. *ACS Photonics* **11**, 2546–2555 (2024).
30. Cao, Z. et al. Aberration-robust monocular passive depth sensing using a meta-imaging camera. *Light.: Sci. Appl.* **13**, 236 (2024).
31. Ji, W. et al. Recent advances in metasurface design and quantum optics applications with machine learning, physics-informed neural networks, and topology optimization methods. *Light Sci. Appl.* **12**, 169 (2023).

32. Ueno, A., Hu, J. & An, S. AI for optical metasurface. *npj Nanophotonics* **1**, 36 (2024).
33. Seo, J. et al. Deep-learning-driven end-to-end metalens imaging. *Adv. Photon.* **6**, 066002 (2019).
34. Liu, K., Wu, J., He, Z. & Cao, L. 4K-DMDNet: diffraction model-driven network for 4K computer-generated holography. *Opto-Electron. Adv.* **6**, 220135–13 (2023).
35. Hao, J. et al. Lensless complex amplitude demodulation based on deep learning in holographic data storage. *Opto-Electron. Adv.* **6**, 220157–15 (2023).
36. Xu, S. et al. Local feature matching using deep learning: A survey. *Inf. Fusion* **107**, 102344 (2024).
37. Liu, X. et al. Edge enhanced depth perception with binocular meta-lens. *Opto-Electron. Sci.* **3**, 230033–230033 (2024).
38. Wang, S. et al. A broadband achromatic metalens in the visible. *Nat. Nanotechnol.* **13**, 227–232 (2018).
39. Xiao, X. et al. Large-scale achromatic flat lens by light frequency-domain coherence optimization. *Light.: Sci. Appl.* **11**, 323 (2022).
40. Yang, L. et al. Depth anything V2. Preprint at http://arxiv.org/abs/arXiv:2406.09414 (2024).

Acknowledgements

We are grateful for financial support by the National Natural Science Foundation of China (62305184); The Major Key Project of PCL (PCL2024A01); Shenzhen Municipal Basic Research (Key Project) (JCY20241202123919027); Basic and Applied Basic Research Foundation of Guangdong Province (2023A1515012932); Science, Technology and Innovation Commission of Shenzhen Municipality (WDZC20220818100259004); the Research Grants Council of the Hong Kong Special Administrative Region, China [Project No. C5031-22G; CityU11310522; CityU11300123]; City University of Hong Kong [Project No. 9610628].

Author contributions

Y.S., Y.Z., M.K.C., and Z.G. conceived the idea for this work. M.K.C., and Z.G. supervised the research. X.L., M.K.C., and Z.G. are responsible for the design of the meta-devices. T.T. fabricated the meta-devices. Y.S., Y.Z., M.K.C., and Z.G. build the measurement system and conduct experimental measurements. Y.S., Y.Z., M.K.C., and Z.G. perform data processing and analysis. All the authors discussed the results and contributed to the preparation of the manuscript and discussions.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44310-025-00070-9.

Correspondence and requests for materials should be addressed to Mu Ku Chen or Zihan Geng.