nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-22696-x OPEN

Neural network based 3D tracking with a graphene transparent focal stack imaging system

Dehui Zhang[1,3], Zhen Xu[1,3], Zhengyu Huang[1,3], Audrey Rose Gutierrez[1], Cameron J. Blocker[1], Che-Hung Liu[1], Miao-Bin Lien[1], Gong Cheng[1], Zhe Liu[1], Il Yong Chun[2✉], Jeffrey A. Fessler[1✉], Zhaohui Zhong[1✉] & Theodore B. Norris[1✉]

Recent years have seen the rapid growth of new approaches to optical imaging, with an emphasis on extracting three-dimensional (3D) information from what is normally a two-dimensional (2D) image capture. Perhaps most importantly, the rise of computational imaging enables both new physical layouts of optical components and new algorithms to be implemented. This paper concerns the convergence of two advances: the development of a transparent focal stack imaging system using graphene photodetector arrays, and the rapid expansion of the capabilities of machine learning including the development of powerful neural networks. This paper demonstrates 3D tracking of point-like objects with multilayer feedforward neural networks and the extension to tracking positions of multi-point objects. Computer simulations further demonstrate how this optical system can track extended objects in 3D, highlighting the promise of combining nanophotonic devices, new optical system designs, and machine learning for new frontiers in 3D imaging.

[1] Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA. [2] Department of Electrical Engineering, University of Hawai'i at Manoa, Honolulu, HI, USA. [3] These authors contributed equally: D. Zhang, Z. Xu, Z. Huang. ✉email: iychun@hawaii.edu; fessler@umich.edu; zzhong@umich.edu; tnorris@umich.edu

Emerging technologies such as autonomous vehicles demand imaging technologies that can capture not only a 2D image but also the 3D spatial position and orientation of objects. Multiple solutions have been proposed, including LiDAR systems[1–3] and light-field cameras[4–7], though existing approaches suffer from significant limitations. For example, LiDAR is constrained by size and cost, and most importantly requires active illumination of the scene using a laser, which poses challenges of its own, including safety. Light-field cameras of various configurations have also been proposed and tested. A common approach uses a microlens array in front of the sensor array of a camera[4,5]; light emitted from the same point with different angles is then mapped to different pixels to create angular information. However, the mapping to a lower dimension carries a tradeoff between spatial and angular resolution. Alternatively, one can use optical masks[6] and camera arrays[7] for light field acquisition. However, the former method sacrifices the signal-to-noise ratio and might need a longer exposure time in compensation; the latter device size could become a limiting factor in developing compact cameras. Single-element position-sensitive detectors, such as recently developed graphene-based detectors[8–10] can provide high speed angular tracking in some applications, but do not provide full 3D information.

To have its highest possible sensitivity to light, a photodetector would ideally absorb all the light incident upon it in the active region of the device. It is possible, however, to design a detector with a photoresponse sufficiently large for a given application, that nevertheless does not absorb all the incident light[11–16]. Indeed, we have shown that a photodetector in which the active region consists of two graphene layers can operate with quite high responsivities, while absorbing only about 5% of the incident light[17]. By fabricating the detector on a transparent substrate, it is possible to obtain responsivities of several A/W while transmitting 80–90% of the incident light, allowing multiple sensor planes to be stacked along the axis of an optical system. We have previously demonstrated a simple 1D ranging application using single pixel of such detectors[18]. We also showed how focal stack imaging is possible in a single exposure if transparent detector arrays can be realized, and developed models showing how light-field imaging and 3D reconstruction could be accomplished.

While the emphasis in ref. [18] was on 4D light field imaging and reconstruction from a focal stack, some optical applications, e.g., ranging and tracking, do not require computationally expensive 4D light field reconstruction[19,20]. The question naturally arises as to whether the focal stack geometry will allow optical sensor data to provide the necessary information for a given application, without reconstructing a 4D light field or estimating a 3D scene structure via depth map. The simple intuition behind the focal stack geometry is that each sensor array will image sharply a specific region of the object space, corresponding to the depth of field for each sensor plane. A stack of sensors thus expands the total system depth of field. The use of sophisticated algorithms, however, may provide useful information even for regions of the object space that are not in precise focus.

The concept of a focal-stack imaging system based on simultaneous imaging at multiple focal planes is shown in Fig. 1a. In the typical imaging process, the camera lens projects an arbitrary object (in this case a ball-and-stick model) onto a set of transparent imaging arrays stacked at different focal planes. With the sensor arrays having a typical transparency on the order of 90%, sufficient light propagates to all planes for sensitive detection of the projected light field. (Of course the final sensor in the stack need not be transparent, and could be a conventional opaque sensor array). Each of the images in the stack records the light distribution at a specific depth, so that depth information is encoded in the image stack. We can then use neural networks to process the 3D focal stack data and estimate the 3D position and configuration of the object.

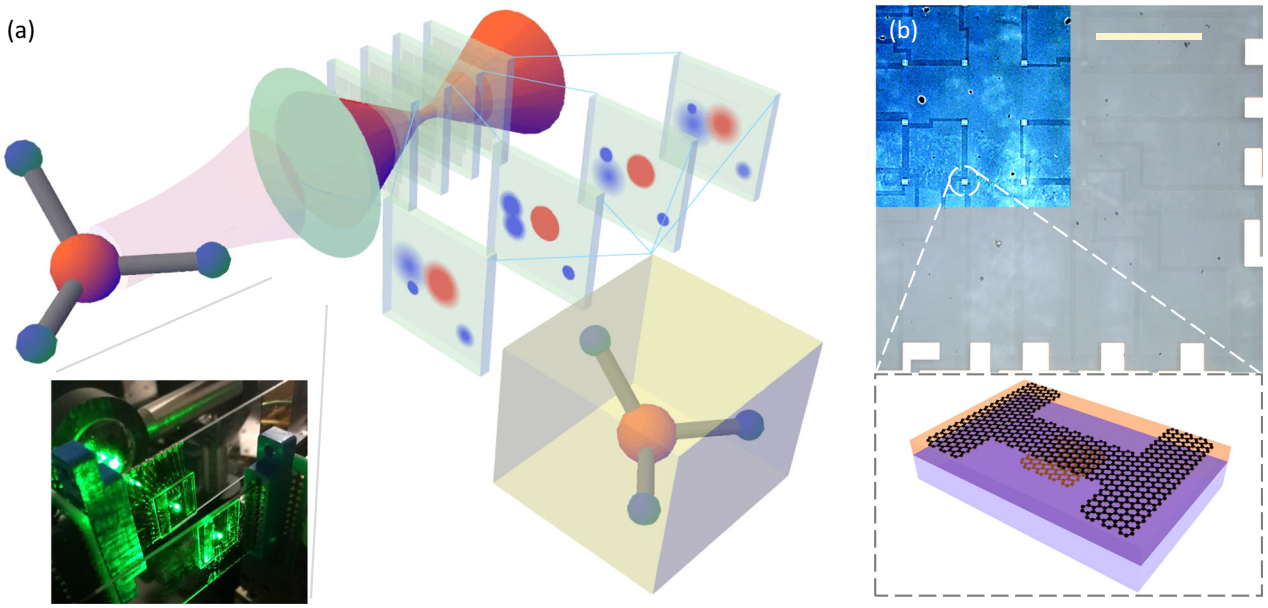


Fig. 1 Concept of focal stack imaging system enabled by focal stacks of transparent all-graphene photodetector arrays. a Schematic showing simultaneous capture of multiple images of a 3D object (ball-and-stick model) on different focal planes. Transparent detector arrays (transparent blue sheets) are placed after the lens (green oval) to form the camera system. The depth information is encoded in the image stacks. Artificial neural networks process the image data and extract important 3D configuration information of the object. Inset: photograph of imaging system used in experiments with two transparent focal planes. **b** Upper panel: Optical image of a 4 × 4 transparent graphene photodetector array, scale bar: 500 μm. Upper-left corner is with false color and enhanced contrast in order to highlight the patterns. Lower panel: Schematic of the all-graphene phototransistor design. It includes a top graphene layer as transistor channel and a bottom graphene patch as floating gate, separated by a 6-nm silicon tunneling barrier (purple). The device is fabricated on transparent glass substrate (blue), and the active detector region is wired out with wider graphene stripes as interconnects.

This work demonstrates a transparent focal stack imaging system that is capable of tracking single and multiple point objects in 3D space, without the need for light field reconstruction. The proof-of-concept experiment is demonstrated with a vertical stack of two 4 × 4 (16-pixel) graphene sensors and feed-forward neural networks that have the form of a multilayer perceptron (MLP)[21]. We also acquired focal stack data sets using a conventional CMOS camera with separate exposures for each focal plane. The simulations demonstrate the capability of future higher-resolution sensor arrays for tracking extended objects. Our experimental results show that the graphene-based transparent photodetector array is a scalable solution for 3D information acquisition, and that a combination of transparent photodetector arrays and machine learning algorithms can lead to a compact camera design capable of capturing real-time 3D information with high resolution. This type of optical system is potentially useful for emerging technologies such as face recognition, autonomous vehicles and unmanned aero vehicle navigation, and biological video-rate 3D microscopy, without the need for an integrated illumination source. Graphene-based transparent photodetectors can detect light with a broad bandwidth from visible to mid-infrared. This enables 3D infrared imaging for even more applications.

Results

All-graphene transparent photodetector arrays. Photodetector arrays with high responsivity and high transparency are central to realizing a focal stack imaging system. To this end, we fabricated all-graphene transparent photodetector arrays as individual sensor planes. Briefly, CVD-grown graphene on copper foil was wet transferred[22] onto a glass substrate and patterned into floating gates of phototransistors using photolithography. We then sputtered 6 nm of undoped silicon on top as a tunneling barrier, followed by another layer of graphene transferred on top and patterned into the interconnects and device channels. (Fig. 1b bottom inset; Details in Supplementary Information 1) In particular, using an atomically thin graphene sheet for the interconnects reduces light scattering when compared to using ITO or other conductive thin films, which is crucial for recording photocurrent signal across all focal stacks. As a proof-of-concept, we fabricated 4 × 4 (16-pixel) transparent graphene photodetector arrays, as shown in Fig. 1b. The active region of each device, the interconnects, and the transparent substrate are clearly differentiated in the optical image due to their differing numbers of graphene layers. The device has an overall raw transparency > 80%; further simulation shows that the transparency can be improved to 96% by refractive index compensation (see Supplementary Information 1). The devices are wired out separately and connected to metal pads, which are then wire-bonded to a customized signal readout circuit. During normal operation, a bias voltage is applied across the graphene channel and the current flowing across the channel is measured; light illumination induces a change in the current, producing photocurrent as the readout (details in Supplementary Fig. 1(a)). The photodetection mechanism of our device is attributed to the photogating effect[17,21,23] in the graphene transistor.

The yield and uniformity of devices were first characterized by measuring the channel conductance. Remarkably, the use of graphene interconnects can still lead to high device yield; 99% of the 192 devices tested show good conductivities (see Supplementary Fig. 1c). The DC photoresponsivity of an individual pixel within the array can reach ~3 A/W at a bias voltage of 0.5 V, which is consistent with the response of single-pixel devices reported previously[18]. We also notice the large device-to-device variation that is intrinsic to most nanoelectronics. Normalization within the array, however, can compensate for this uniformity issue, which is a common practice even in a commercial CCD array.

To reduce the noise and minimize device hysteresis, the AC photocurrent of each pixel is recorded for 3D tracking and imaging. This measurement scheme sacrifices responsivity but makes the measurement faster and more reliable. As shown in Fig. 2a, a chopper modulates the light and a lock-in amplifier records the AC current at the chopper frequency. The power dependence of the AC photocurrent is also examined (see Supplementary Fig. 1e). The responsivity remains constant in the power range that we use to perform our test. Hence only a single exposure is required to calibrate the nonuniformity between the pixels. We note that the graphene detector speed is currently limited by the charge traps within the sputtered silicon tunneling barrier[17], which can be improved through better deposition techniques and design, as well as higher quality materials[24].

Focal stack imaging with transparent sensors. The concept of focal stack imaging was demonstrated using two vertically stacked transparent graphene arrays. As shown in Fig. 2a, two 4 × 4 sensor arrays were mounted vertically along the optical axis, separated at a controlled distance, to form a stack of imaging planes. This double-focal-plane stack essentially serves as the camera sensor of the imaging system. A convex lens focuses a 532 nm laser beam, with the beam focus serving as a point object. The focusing lens was mounted on a 3D-motorized stage to vary the position of the point object in 3D. The AC photocurrent is recorded for individual pixels on both front and back detector arrays while the point object is moving along the optical axis.

Figure 2b shows a representative set of images captured experimentally by the two detector arrays when a point object is scanned at different positions along the optical axis (12 mm, 18 mm, 22 mm) respectively, corresponding to focus shifting from the back plane toward the front plane (Fig. 2c). The grayscale images show the normalized photoresponse, with white (black) color representing high (low) intensity. As the focus point shifts from the back plane toward the front plane, the image captured by the front plane shrinks and sharpens, while the image captured by the back plane expands and blurs. Even though the low pixel density limits the image resolution, these results nevertheless verify the validity of simultaneously capturing images at multiple focal planes.

3D tracking of point objects. While a single image measures the lateral position of objects as in conventional cameras, differences between images captured in different sensor planes contain the depth information of the point object. Hence focal stack data can be used to reconstruct the 3D position of the point object. Here we consider three different types of point objects: a single-point object, a three-point object, and a two-point object that is rotated and translated in three dimensions.

First, we consider single-point tracking. In this experiment, we scanned the point source (dotted circle in Fig. 2a) in a 3D spatial grid of size 0.6 × 0.6 mm (x, y axes) × 20 mm (z axis, i.e., the longitudinal direction). The grid spacing was 0.06 mm along the x, y axes, and 2 mm along the z axis, leading to 1331 grid points in total. For each measurement, two images were recorded from the graphene sensor planes. We randomly split the data into two subsets, training data with 1131 samples (85% of total samples) and testing data with 200 samples (15% of total samples); all experiments used this data splitting procedure. To estimate three spatial coordinates of the point object from the focal stack data, we trained three separate MLP[25] neural networks (one for each spatial dimension) with mean-square error (MSE) loss. The

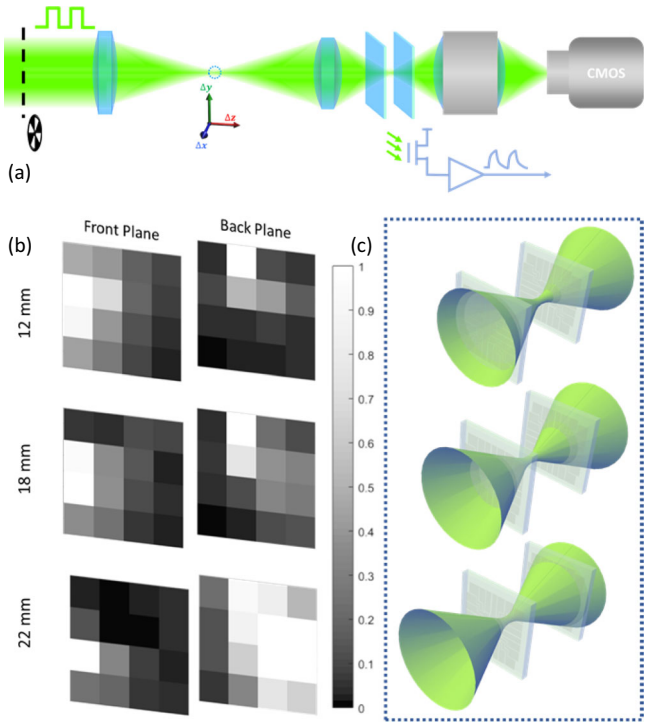


Fig. 2 Experimental demonstration of focal stack imaging using double stacks of graphene detector arrays. a A schematic of measurement setup. A point object (dotted circle) is generated by focusing a green laser beam (532 nm) with the lens. Its position is controlled by a 3D motorized stage. Two detector arrays (blue sheets) are placed behind the lens. An objective and CMOS camera are placed behind the detector array for sample alignment. A chopper modulates the light at 500 Hz and a lock-in amplifier records the AC current at the chopper frequency. **b** Images captured by the front and back photodetector planes with objects at three different positions along the optical axis (12 mm, 18 mm, 22 mm, respectively). The grayscale images are generated using responsivities for individual pixels within the array, normalized by the maximum value for better contrast. The point source is slightly off-axis in the image presented, leading to the shift of spot center. **c** The illustrations of the beam profiles corresponding to the imaging planes in (**b**). The focus is shifting from the back plane (top panel) toward the front plane (bottom panel).

results (Fig. 3a, b) show that even with the limited resolution provided by 4×4 arrays, and only two sensor planes, the point object positions can be determined very accurately. We used the root-mean-square error (RMSE) to quantify the estimation accuracy on the testing dataset; we obtained RMSE values of 0.012 mm, 0.014 mm, and 1.196 mm along the x, y, and z directions, respectively.

Given the good tracking performance with the small-scale (i.e., 4×4 arrays) graphene transistor focal stack, we studied how the tracking performance scales with array size. We determined the performance advantages of larger arrays by using conventional CMOS sensors to acquire the focal stack data. For each point source position, we obtained multi-focal plane image stacks by multiple exposures with varying CMOS sensor depth (note that focal stack data collected by CMOS sensors with multiple exposures would be comparable to that obtained by the proposed transparent array with a single exposure, as long as the scene being imaged is static), and down-sampled the resolution of high resolution (1280×1024) images captured by CMOS sensor to 4×4, 9×9, and 32×32. We observed that tracking performance improves as the array size increases; results are presented in Supplementary Table 1.

We next considered the possibility of tracking multi-point objects. Here, the object consisted of three point objects, and these three points can have three possible relative positions to each other. We synthesized 1880 3-point objects images as the sum of single-point objects images from either the graphene detectors or the CMOS detectors (see details of focal stack synthesis in Supplementary Information 2). This synthesis approach is reasonable given that the detector response is sufficiently linear and it avoids the complexity of precisely positioning multiple point objects in the optical setup. To estimate the spatial coordinates of the 3-point synthetic objects, we trained an MLP neural network with MSE loss that considers the ordering ambiguity of the network outputs (see Supplementary Information 2, Equation (1)). We used 3-point object's data synthesized from the CMOS-sensor readout in the single-point tracking experiment (with each CMOS image smoothed by spatial averaging and then down-sampled to 9×9). We found that the trained MLP neural network can estimate a multi-point object's position with remarkable accuracy; see Fig. 3c, d. The RMSE values calculated from the entire test set are 0.017 mm, 0.016 mm, 0.59 mm, along x-, y-, z-directions, respectively. Similar to the single-point object tracking experiment, the multi-point object tracking performance improves with increasing sensor resolution (see Supplementary Tables 2–4).

Finally, we considered tracking of a two-point object that is rotated and translated in three dimensions. This task aims to demonstrate 3D tracking of a continuously moving object, such as a rotating solid rod. Similar to the 3-point object tracking experiment, we synthesized a 2-point object focal stack from single-point object focal stacks captured using the graphene transparent transistor array. The two points are located at the same x-y plane and are separated by a fixed distance, as if tied by a solid rod. The rod is allowed to rotate in the x-y plane and translate along the z-axis, forming helical trajectories, as shown in Fig. 3e. We trained an MLP neural network with 242 training trajectories using MSE loss to estimate the object's spatial coordinates and tested its performance on 38 test rotating trajectories. Figure 3e shows the results of one test trajectory. The neural network estimated the orientation (x- and y-coordinates) and depth (z-coordinate) of test objects with good accuracy: RMSE along x-, y-, and z-directions for the entire test set are 0.016 mm, 0.024 mm, 0.65 mm, respectively.

Supplementary Information 2 gives further details on the MLP neural network architectures and training.

3D extended object tracking. The aforementioned objects consisted of a few point sources. For non-point-like (extended) objects, the graphene 4×4 pixel array fails to accurately estimate the configuration, given the limited information available from such a small array. To illustrate the possibilities of 3D tracking of a complex object and estimating its orientation, we used a ladybug as an extended object and moved it in a 3D spatial grid of size $8.5 \times 8.5 \times 45$ mm. The grid spacing was 0.85 mm along both x- and y-directions, and 3 mm along z-direction. At each grid point, the object took 8 possible orientations in the x-z plane, with 45° angular separation between neighboring orientations (see experiment details in Supplementary Information 2). We acquired 15,488 high-resolution focal stack images using the CMOS sensor (at two different planes) and trained two convolutional neural networks (CNNs), one to estimate the ladybug's position and the other for estimating its orientation, with MSE loss and the cross-entropy loss, respectively. Figure 4 shows the results for five test samples. The CNNs correctly classified the orientation of all five samples and estimated their 3D position accurately. For the entire test set, the RMSE along x-, y-, and

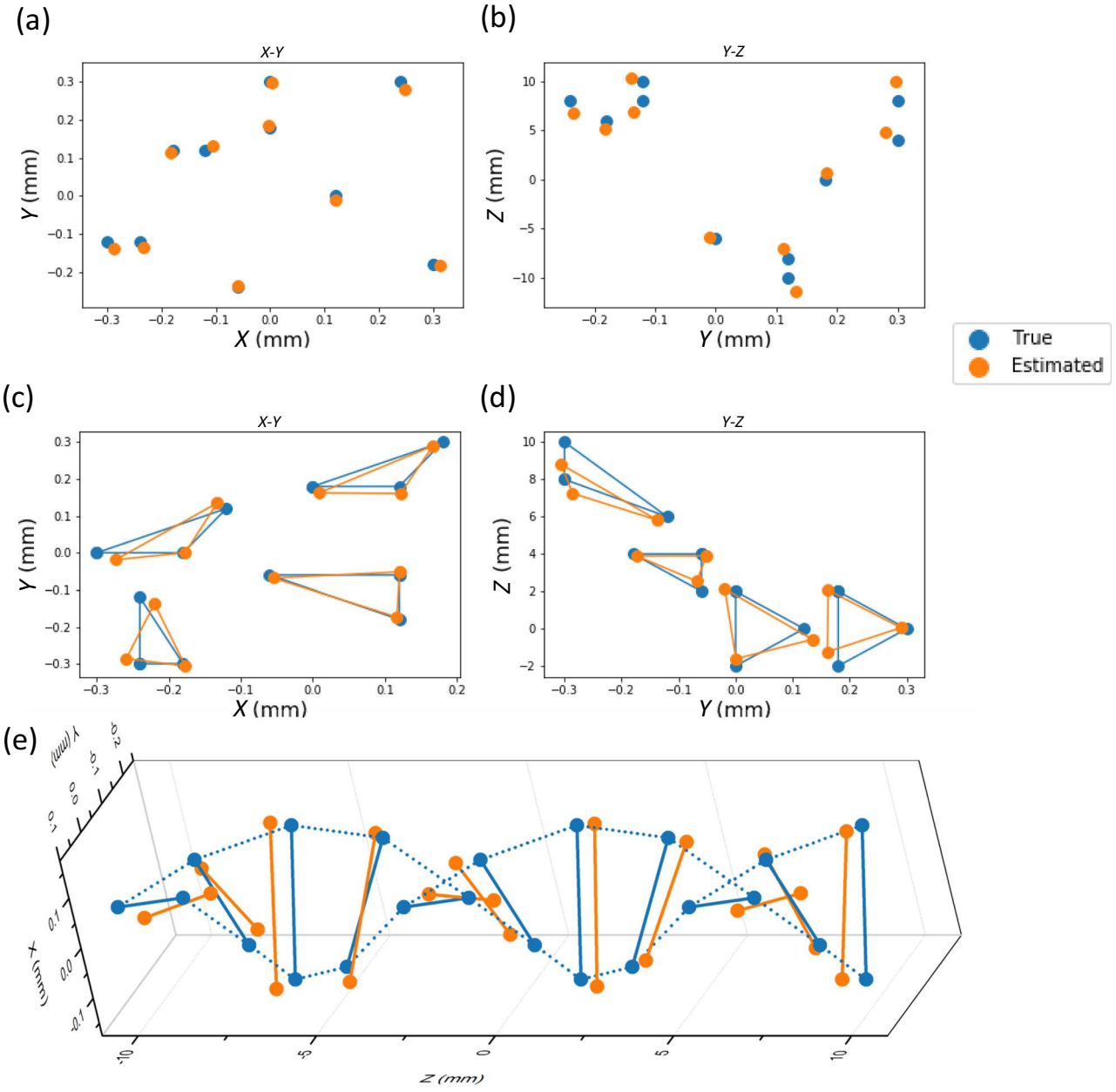


Fig. 3 3D point object tracking using focal stack data for three different types of point objects. a, **b** Tracking results for single point object (only 10 test samples are shown). Results are based on images captured with the graphene photodetector arrays. **c**, **d** Tracking results for three-points objects (only 4 test samples are shown). Results are based on data synthesized from multi-focal-plane CMOS images (downsampled to 9×9) of single point source. **e** Tracking results for rotating two-point objects on one testing trajectory. The object is rotating counter-clockwise (viewed from left) while moving from $z = -10$mm to $z = 10$mm. Results are based on data synthesized from single point source images captured with graphene photodetector arrays.

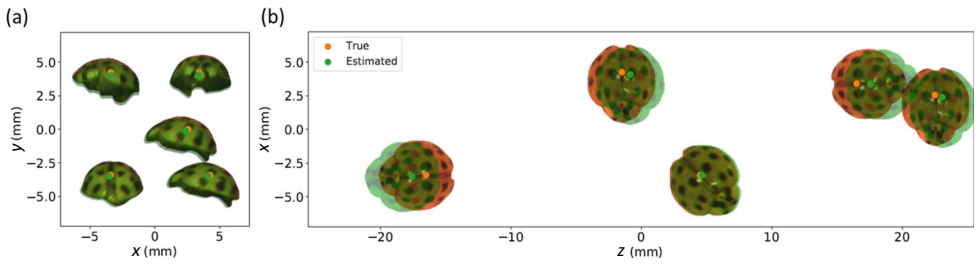


Fig. 4 3D extended-object tracking and its orientation estimation using focal stack data collected by a CMOS camera. (**a**) Results in the *x*-*y*-plane perspective and (**b**) in the *x*-*z*-plane perspective. The estimated (true) ladybug's position and orientation are indicated by green (orange) dots and green (orange) overlaid ladybug images. Note that the ladybug images are not a part of the neural network output and are shown for illustration only.

z-directions is 0.11 mm, 0.13 mm, and 0.65 mm, respectively, and the orientation is classified with 99.35% accuracy. We note that at least two imaging planes are needed to achieve good estimation accuracy along depth (z)-direction: when the sensor at the front position is solely used, the RMSE value along z-direction is 2.14 mm, and when the sensor at the back position is solely used, the RMSE value along z-direction is 1.60 mm.

Supplementary Fig. 6 describes the CNN architectures and training details.

Discussion

In conclusion, we designed and demonstrated a focal stack imaging system enabled by graphene transparent photodetector arrays and the use of feedforward neural networks. Even with limited pixel density, we successfully demonstrated simultaneous imaging at multiple focal planes, which can be used for 3D tracking of point objects with high speed and high accuracy. Our computer model further proves that such an imaging system has the potential to track an extended object and estimate its orientation at the same time. Future advancements in graphene detector technology, such as higher density arrays and smaller hysteresis enabled by higher quality tunnel barriers, will be necessary to move beyond the current proof-of-concept demonstration. We also want to emphasize that the proposed focal stacking imaging concept is not limited to graphene detectors alone. Transparent (or semi-transparent) detectors made from other 2D semiconductors and ultra-thin semiconductor films can also be implemented as the transparent sensor planes within the focal stacks. The resulting ultra-compact, high-resolution, and fast 3D object detection technology can be advantageous over existing technologies such as LiDAR and light-field cameras. Our work also showcases that the combination of nanophotonic devices, which is intrinsically high-performance but non-deterministic, with machine learning algorithms can complement and open new frontiers in computational imaging.

Data availability

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

Code availability

The code is accessible at https://zenodo.org/record/4282790#.X7gKPshKguU.

Received: 5 August 2020; Accepted: 26 February 2021;
Published online: 23 April 2021

References

1. Schwarz, B. LIDAR: mapping the world in 3D. *Nat. Photonics* **4**, 429 (2010).
2. Oggier, T., Scott T. S. & A. Herrington. Line scan depth sensor. U.S. Patent Application No. 15/700,231.
3. Niclass, C. L. et al. Light detection and ranging sensor. U.S. Patent Application No. 15/372,411.
4. Ng, R. et al. Light field photography with a hand-held plenoptic camera. *Computer Sci. Tech. Rep. CSTR2.* **11**, 1–11 (2005).
5. Navarro, H. et al. High-resolution far-field integral-imaging camera by double snapshot. *Opt. Express* **20**, 890–895 (2012).
6. Xu, Z., Ke, J. & Lam, E. Y. High-resolution lightfield photography using two masks. *Opt. Express* **20**, 10971–10983 (2012).
7. Venkataraman, K. et al. PiCam: an ultra-thin high performance monolithic camera array. *ACM Trans. Graph. (TOG)* **32**, 166 (2013).
8. Wang, W. et al. High-performance position-sensitive detector based on graphene–silicon heterojunction. *Optica* **5**, 27–31 (2018).
9. Liu, K. et al. Graphene-based infrared position-sensitive detector for precise measurements and high-speed trajectory tracking. *Nano Lett.* **19**, 8132–8137 (2019).
10. Wang, W.-H. et al. Interfacial amplification for graphene-based position-sensitive-detectors. *Light.: Sci. Appl.* **6**, e17113 (2017).
11. Liu, N. et al. Large-area, transparent, and flexible infrared photodetector fabricated using PN junctions formed by N-doping chemical vapor deposition grown graphene. *Nano Lett.* **14**, 3702–3708 (2014).
12. Zheng, Z. et al. Flexible, transparent and ultra-broadband photodetector based on large-area WSe2 film for wearable devices. *Nanotechnology* **27**, 225501 (2016).
13. Tsai, S.-Y. I., Hon, M.-H. & Lu, Y.-M. Fabrication of transparent p-NiO/n-ZnO heterojunction devices for ultraviolet photodetectors. *Solid-State Electron.* **63**, 37–41 (2011).
14. Tanaka, H. et al. Transparent image sensors using an organic multilayer photodiode. *Adv. Mater.* **18**, 2230–2233 (2006).
15. Stiebig, H. et al. Standing wave detection by thin transparent n–i–p diodes of amorphous silicon. *Thin Solid Films* **427**, 152–156 (2003).
16. Jovanov, V. et al. Transparent fourier transform spectrometer. *Opt. Lett.* **36**, 274–276 (2011).
17. Liu, C. H. et al. Graphene photodetectors with ultra-broadband and high responsivity at room temperature. *Nat. Nanotechnol.* **9**, 273–278 (2014).
18. Lien, M. B. et al. Ranging and light field imaging with transparent photodetectors. *Nat. Photonics* **14**, 143–148 (2020).
19. Blocker C. J., Chun Il Y, & Fessler J. A. Low-rank plus sparse tensor mod- els for light-field reconstruction from focal stack data. In Proc. IEEE Image, Video, and Multidim. Signal Process. (IVMSP) Workshop, pp. 1–5, Zagori, Greece, Apr. 2018.
20. Chun, I. Y. et al. Momentum-Net: Fast and convergent iterative neural network for inverse problems. *IEEE. Trans. Pattern. Anal. Mach. Intell.* (2020).
21. Konstantatos, G. et al. Hybrid graphene–quantum dot phototransistors with ultrahigh gain. *Nat. Nanotechnol.* **7**, 363 (2012).
22. Lee, S. et al. Homogeneous bilayer graphene film based flexible transparent conductor. *Nanoscale* **4**, 639–644 (2012).
23. Sun, Z. et al. Infrared photodetectors based on CVD-grown graphene and PbS quantum dots with ultrahigh responsivity. *Adv. Mater.* **24**, 5878–5883 (2012).
24. Zhang, D. et al. Electrically tunable photoresponse in a graphene heterostructure photodetector[C]//2017 Conference on Lasers and Electro-Optics (CLEO). IEEE, 2017: 1–2.
25. Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Rev.* **65**, 386 (1958).

Acknowledgements

We gratefully acknowledge financial support from the W. M. Keck Foundation and National Science Foundation grants IIS 1838179. Devices were fabricated in the Lurie Nanofabrication Facility at University of Michigan, a member of the National Nanotechnology Infrastructure Network funded by the National Science Foundation.

Author contributions

D.Z., Z.X., Z.H., J.A.F., Z.Z., and T.B.N. conceived the experiments. D.Z. and Z.L. fabricated the devices. Z.X., D.Z., C.L., M.L., and G.C. built the optical setup. D.Z., Z.X., and A.R.G. performed the nanodevice optoelectrical measurements. Z.H. performed the CMOS camera data collection. Z.H., I.Y.C., and C.J.B. worked on neural network based 3D reconstructions. All authors discussed the results and co-wrote the manuscript.

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/s41467-021-22696-x.

Correspondence and requests for materials should be addressed to I.Y.C., J.A.F., Z.Z. or T.B.N.

Peer review information *Nature Communications* thanks David Brady, Yang Xu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary Information

Supplementary Information 1: System Hardware

I. Graphene Transparent Photodetector Device Fabrication:

We first transferred a layer of graphene onto a commercial glass substrate. The graphene layer was grown on copper foil with chemical vapor deposition (CVD), and a standard wet transfer process produced a decent coverage of the monolayer graphene on the centimeter scale[1]. We patterned the graphene layer into isolated squares (as the floating gate) using photolithography. Then we etched away the exposed graphene with oxygen plasma. A 6-nm layer of sputtered silicon was sputtered on top of the graphene layer as the tunneling barrier. Another layer of graphene was transferred on top of the barrier immediately after silicon sputtering to minimize surface oxidation. We annealed the sample in Ar at 300 C for 15 mins to enhance graphene's adhesion with the substrate, so that there is less stripping-off in the subsequent process. This graphene layer was lithographically patterned into the channel of the phototransistor as well as interconnects. The individual device pixels were spaced 0.3 mm away from each other. Cr/Au metal contacts were then deposited and connected to the graphene interconnects, leaving a 2.5 mm by 2.5 mm transparent window for light to pass through. The Cr/Au pads were then wire bonded to a sample holder in the readout circuit. The readout circuit was a scanning line that selectively applies bias to different pixels and collects the photocurrent in the target pixel.

II. Graphene Photodetector Characterization:

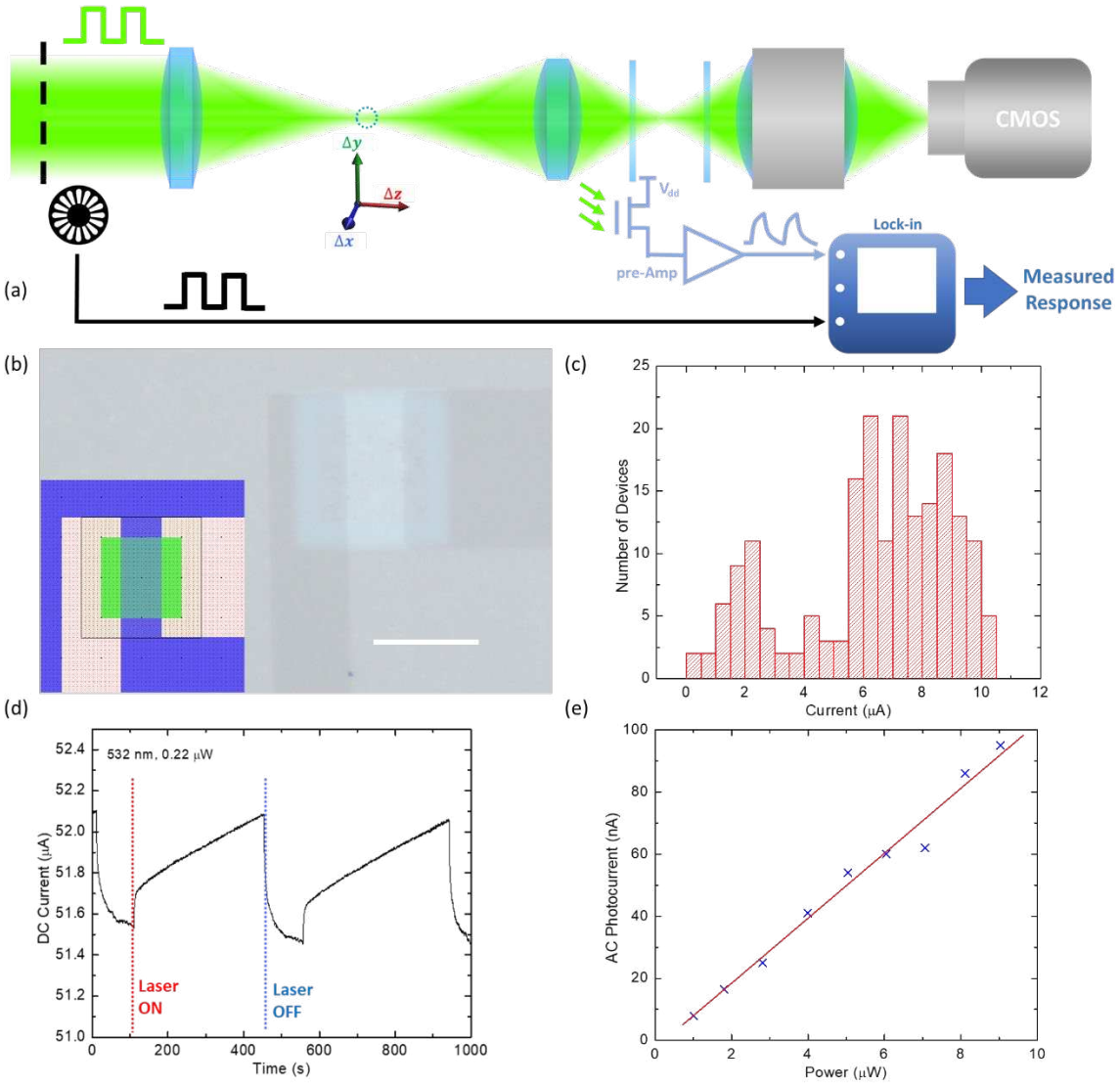
The fabricated all-graphene devices in general showed good coverage over the entire die. To check the uniformity and device yield, we applied a 0.5 V bias voltage across every pixel and measured the current across the graphene channel (see Supplementary Fig. 1(c)). Of all 192 devices tested, only 2 devices showed open circuit. This DC conductance test result shows 99% graphene device yield and negligible graphene peeling-off during the nanofabrication process.

The detector photoresponse is characterized using a 532 nm Verdi V10 CW laser under 0.22 μW illumination on a single pixel. For the responsivity calibration of both detector layers, we first align the lateral position of the imaging chip to maximize the photocurrent readout from the center pixel. This ensures that the optical beam is centered on the chip. Then we move the lens of the camera system along

the optical axis to provide a beam spot with a diameter > 4 mm. The large spot size provides nearly uniform illumination in the 0.9-mm-wide detector array. The beam sizes are measured using a power meter and a blade as the moving mask. Then we measure the photocurrent from the array. By calculating the illumination power per device area, we calculate the responsivities of the devices.

The responsivity is 3 A/W under 0.5 V source-drain bias. It is slightly smaller than previously reported values due to a relatively small bias voltage applied, lower doping level of graphene, and the geometry of the device. The DC photocurrent shows both a fast response on the scale of seconds and a slow response in hundreds of seconds, which is due to the charge trapping effect of the highly defective silicon barrier. By changing the dielectric material, the response time of such a structure can be decreased to the sub-millisecond level[2].To increase the speed of measurement and remove effects from drifting dark currents, we adopted an AC photocurrent measurement scheme to measure the smaller but fast component, as discussed in the main text. Supplementary Fig. 1(e) shows a linear dependence of AC photocurrent with respect to illumination power. A linear power dependence of the AC photocurrent was observed.

We also characterize the transparency of the graphene detector array with a light focused on the detector plane. Transmission of a 532 nm laser beam through the array is measured to be 81%, while the reflection of the uncoated glass substrate contributes to a transmission of 86%, as measured in the graphene-free area of the device. The graphene detector array contributes to only 5% of decrease in transmission. If the application requires, the transparency can be significantly improved to >95% using an antireflection coating, and by replacing the silicon layer with ALD-grown Al_2O_3 of the same thickness.

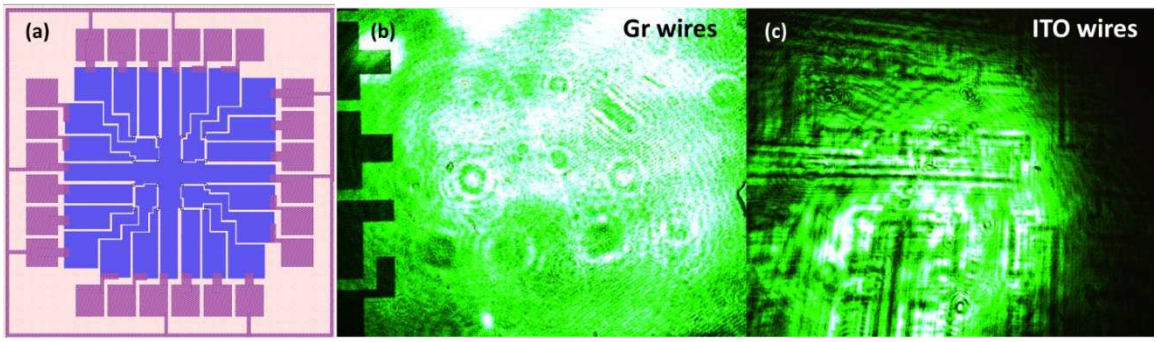


Supplementary Figure 1. Optical and electrical measurements on the all-graphene transparent photodetector. (a) measurement setup. A point object (dotted circle) is generated by focusing a green laser beam (532 nm) with the lens. Its position is controlled by a 3D motorized stage. Two detector arrays (blue sheets) are placed behind the lens. An objective and CCD camera are placed behind the detector array for sample alignment. A chopper modulates the light at 500 Hz and a lock-in amplifier records the AC current at the chopper frequency. (b) Optical microscope image and layout diagram (inset) of a single pixel. Blue: top layer graphene channel; green: bottom layer floating gate. The overlapped channel region (separated by the tunneling barrier) is 30 μm by 10 μm. The lower floating gate layer is 20 μm by 20 μm, intentionally made larger to avoid peeling-off. Scale bar: 20 μm. (c) Histogram of the DC currents across graphene channels for individual detector devices. Bias voltage applied is 0.5 V. (d) DC temporal photoresponse of a typical graphene detector following light illumination. Both a fast and a slow

component were observed, while the background current also showed drifting over time. (e) Power dependence of AC photocurrent, which measures the fast component and suppress the background drift. A linear power dependence (red line) is observed.

III. Use of Graphene for Interconnects

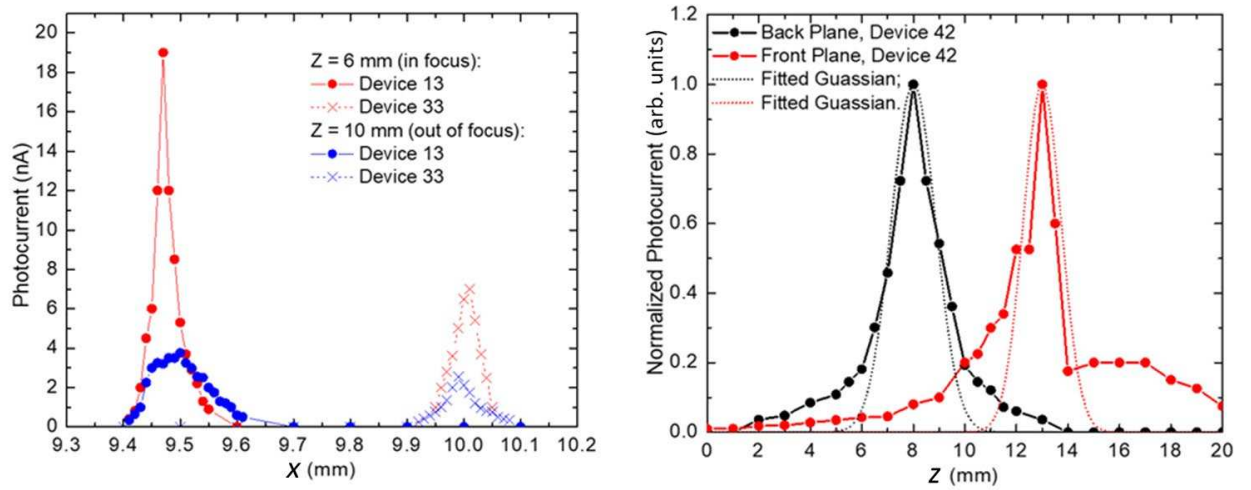
In our transparent detector array design, graphene is used not only as the active pixel material but also as the passive interconnect. Compared with other transparent electrode materials, such as indium-tin oxide (ITO), graphene is atomically thin while maintaining similar conductivity. This ultimate thin-ness minimizes optical interference patterns generated from the interconnect patterning, and also suppresses edge scatterings from normal metal wires. Supplementary Fig. 2 compares the optical transmission images of detector arrays fabricated using graphene interconnects versus ITO interconnects under 532 nm laser illumination. Even though the interference and scattering effect from ITO interconnects can be reduced with a refractive index compensation layer, this would add more complexity to the sensor array.



Supplementary Figure 2. Array design (a) and optical images of photodetector arrays captured by CMOS camera. Samples using graphene interconnects (b) showed significantly weaker effects of interference and scattering than samples with ITO interconnects (c).

IV. Readout of Single Pixel in 3D Ranging

Before using the graphene devices as an imaging array, an imaging hardware reliability test is performed (Supplementary Fig. 3). The single-pixel photocurrent is measured while the light source moves both in-plane and along the optical axis. For each pixel tested, the photocurrent always shows a single peak when the object is translated in 3 dimensions. The peak positions and FWHMs of peaks match the relative positions of pixels tested, confirming that the graphene detector array can indeed accurately “image” the focus of a point light source.



Supplementary Figure 3. 3D ranging test of single detector pixels. Left: Photocurrents from two pixels (Device 13 and 33) spaced laterally in the *X* direction. When the point light source moves in *X* direction, the devices are illuminated sequentially and give peaks at two different *X* positions. When the point source is off focus on the device plane, a broader peak is observed corresponding to de-focusing. Right: Photocurrents from two pixels at front and back planes along the optical path (*Z* direction), respectively. As the point source moves from front to back along *Z* direction, the focus shifts from front to back detector plane accordingly, resulting in the observed photocurrent peaks at different *Z* positions. Deviation from the Gaussian fit stems from imperfect beam quality and possible off-axis alignment.

V. Noise Analysis

The noise equivalent power (NEP) is a good measure to discuss the SNR in realistic applications. The NEP of the device has been discussed in the supporting information of our previous work (Liu C H, et al. Nature Nanotechnology, 2014, 9(4): 273-278.). We collect our data with a modulation frequency of 500 Hz. At this frequency, the noise spectral density is 10^{-9} A/Hz$^{1/2}$. The noise level is consistent with the 1/*f* noise of graphene transistors observed[3]. This indicates that the channel's 1/*f* noise dominates over the shot noise of dark currents in the tunneling barrier. With an AC responsivity of 10 mA/W (Supplementary Fig. 1 (e)), the NEP is 0.1 μW/Hz$^{1/2}$. The value is small compared with our test illumination power of ~10 μW per device.

We can also compare this with realistic illumination powers in a camera system. Assume a camera system with a 20-mm aperture and a numerical aperture of 0.7. When using it to image a white Lambertian

surface under sunlight, the estimated optical power per pixel is 0.05 μW. This indicates a relatively low SNR for our current device.

The low SNR is largely due to the slow response of our photodetectors, which is caused by the large density of charge traps in the tunneling barrier. Charge traps capture the tunneling charges and compensate the local field that motivates more interlayer hopping. One of our previous work replaced amorphous silicon with high quality Al_2O_3. The responsivity at 1 kHz is as high as 60 A/W at 532 nm^2. Taking all the corresponding design variations, including increased noise due to a larger channel current, we expect a NEP of 0.1 $nW/Hz^{1/2}$, which is more than enough for realistic applications. In this experiment, we did not adopt the Al_2O_3 barrier due to fabrication yield considerations, as the thin material is vulnerable to the base used in lithography. Nevertheless, there are no fundamental limitations that prevent us from fabricating transparent devices with higher speed and responsivity.

In the above discussion, experimental results suggest that the tunneling current is not the major contribution of noise. For a more complete discussion, we can further analyze the tunneling noise's order of magnitude. The shot noise's current spectral density is $S = 2eI$ when the interlayer bias $V >> kT/e$ [4]. The current is the total of the dark current and the photocurrent, which is around 10 pA in our device. Hence the noise current density of the tunneling photodiode (before amplification) is around 1.8 $fA/Hz^{1/2}$. The value is much smaller than the photocurrent at any realistic illumination power. Also notice that $1/f$ noise is not considered here, so that the estimation only sets the lower limit for the noise amplitude contributed by the tunneling current before amplified with the photogating effect.

Moreover, neural networks can be trained to be robust against input noises. This further lifts the SNR requirements for the reported application.

In conclusion, the device's noise is dominated by the $1/f$ noise in the channel. The photoconductive gain amplifies the noise from the vertical tunneling diode. However, it does not dominate the device noise based on both tests and order-of-magnitude estimation. Better implementation of the device to image ambient objects needs an increase in responsivity. One promising way is to improve the tunneling barrier quality, which is also supported by previous work.

Supplementary Information 2: Data Processing and Machine Learning

I. Single-Point Object Focal Stack from CMOS Camera

We recorded 1,331 single-point object focal stacks using the transparent graphene transistor array and separately using a CMOS sensor (Thorlab DCC1645C); see the right part of the Fig. 2(a) in the main text.

By moving the CMOS sensor along z to focus either closer to or farther away from the lens, we captured focal stacks from CMOS camera. This data allows us to test how the image resolution and image quality of the graphene sensors affect the 3D ranging performance of a machine-learning algorithm.

We applied the following procedure to each high-resolution (1280 × 1024) color image captured by the CMOS camera: we convert the captured color image to gray image and optionally smooth it by spatial averaging and generate low resolution single-point object focal stacks of spatial size 4 × 4, 9 × 9 or 32 × 32. We used the processed images in either single-point tracking (to investigate the effects of imaging resolution to the tracking performance) or synthesizing multi-point object focal stacks.

II. Synthesizing Multi-Point Object Focal Stack

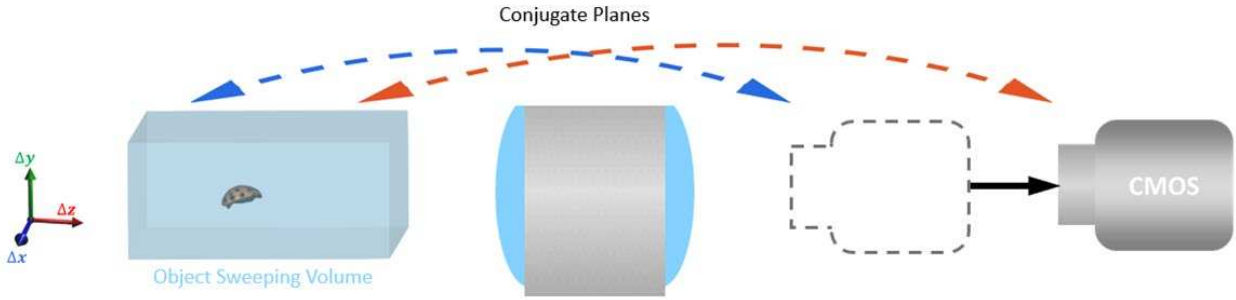
We synthesized multi-point object focal stacks by combining focal stacks from the scanned single point object (either from transparent graphene transistor array or from CMOS camera). The synthesis assumes that the detector's response is linear, i.e., suppose I_i is the sensor image of the single point object at location (x_i, y_i, z_i). Then the sensor image I_{multi} consisting of multiple points is synthesized as $I_{\text{multi}} = \sum_{i=1}^{N} I_i$, where N is the number of point objects.

We constructed an M-point object focal stack dataset, where the dataset consists of multiple subsets, and each subset consists of K possible shapes (relative position between points), by synthesizing each shape independently and then combining them. We translated an object to all possible locations (i.e., no point of M-point object is off the 3D grid) in the 3D $11 \times 11 \times 11$ scanning grid; at each location, we synthesize the corresponding focal stack according to the summation above. The number of synthesized datasets with (M = 2, K = 2), (M = 2, K = 3), (M = 3, K = 2), (M = 3, K = 3) were 1600, 2320, 1232, and 1880, respectively.

We constructed the rotating 2-point object focal stack dataset by selecting focal stacks from the M-point focal stack with K possible shapes dataset, with $M = 2, K = 4$. Four shapes of a 2-point object (i.e., $M = 2, K = 4$) were chosen to have same inter-point distance but rotated by different angles (26.5°, 63.5°, 116.5°, and 153.5°) about z axis (e.g., (1,0,0) means 0° rotation about $z$axis and (0,1,0) means 90° rotation about $z$axis). To form the helical trajectory in the $M = 2, K = 4$ setup, we selected an angle from the set {26.5°, 63.5°, 116.5°, and 153.5°} at each z position in the following sequence: 63.5°, 26.5°,153.5°, 116.5°, 63.5°, 26.5°, 153.5°,116.5°, 63.5°, 26.5°, 153.5°, for z =-10 mm, -8 mm, ..., 10 mm. See graphical illustration in Fig. 3(e) of the main paper.

III. Extended Object Focal Stack

We captured extended object focal stacks using the CMOS sensor. The experimental setup is shown in Supplementary Fig. 4. We used a ladybug as the extended object and moved it in a 3D spatial grid of size 8.5 mm × 8.5 mm × 45 mm. The grid spacing is 0.85 mm along both x and y, 3 mm along z. At each grid point, the object has 8 possible orientations in the x-z plane, with 45° angular separation between neighboring orientations. This led to a total of 1,5488 focal stacks, where each focal stack consists of two images captured by the CMOS sensor positioned at different z positions. Similar to Part B-III, all images are converted to gray images before feeding to the neural networks.



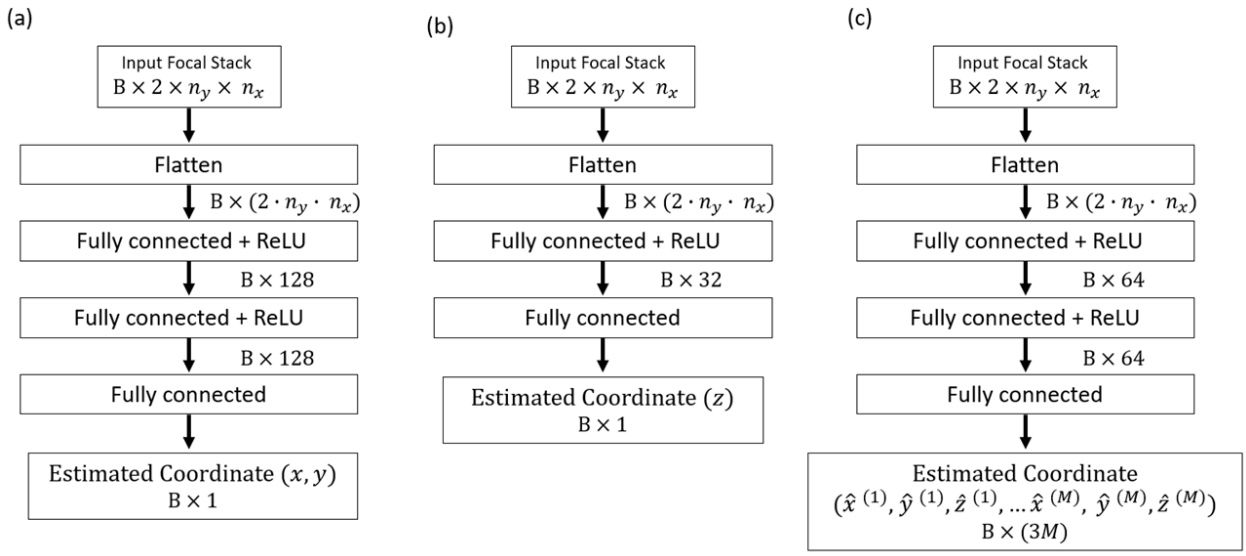
Supplementary Figure 4. Experimental set-up for capturing the extended object (ladybug) focal stack, using CMOS sensor.

IV. Neural Network Architectures and Training

We implemented all neural networks in Pytorch (ver. 1.0). The network architectures and training details are described below.

For single-point object tracking, separate neural networks were trained for estimating the three spatial coordinates x, y and z, respectively. Supplementary Fig. 5(a) shows the network architecture used for estimating coordinates x and y, and Supplementary Fig. 5(b) shows the network architecture used for estimating z. For multi-point object tracking, a single neural network (Supplementary Fig. 5(c)) is trained to estimate all points' coordinates.

In point object tracking cases (Supplementary Fig.5 (a-c)), the focal stack data is flattened into a one-dimensional vector and subsequently passed through multilayer perceptron (MLP)[5] using Rectified Linear Unit (ReLU) as the activation function.



Supplementary Figure 5. Neural network architectures for 3D ranging. B is the general batch size of the data (e.g., in training, B is the training batch size; in testing with a single sample, $B = 1$). (a) Network for estimating single point object's x or y coordinate. (b) Network for estimating single point object's z coordinate. (c) Network for estimating M-point object's (x_i, y_i, z_i) coordinates tuple.

For single-point object tracking, the network outputs a single coordinate value for each focal stack, and the networks are trained by minimizing the following mean-square error (MSE) loss

$$\frac{1}{N}\sum_{i=1}^{N}\left(\hat{s_i} - s_i\right)^2,$$

where N is the number of training samples, s_i is the true spatial coordinate (x_i, y_i, or z_i) and $\hat{s_i}$ is the estimated spatial coordinate from a neural network. We trained networks using the Adam[6] optimizer with the learning rate of 10^{-2}, the training batch size of 50, and 2000 epochs.

For training multi-point object tracking neural networks, we defined the following MSE loss that considers the ordering ambiguity of the network outputs in training:

$$\frac{1}{N}\sum_{i=1}^{N}\min_{(p_1,\ldots,p_M)\in P}\sum_{j=1}^{M}\left(\hat{x}_i^{(j)} - x_i^{(p_j)}\right)^2 + \left(\hat{y}_i^{(j)} - y_i^{(p_j)}\right)^2 + \left(\hat{z}_i^{(j)} - z_i^{(p_j)}\right)^2, \quad (1)$$

where M is the number of points of the object, P is the set containing all possible permutations of the tuple $(1,2,\dots M)$, $x_i^{(j)}$ and $\hat{x}_i^{(j)}$ are the true and estimated coordinate of the i^{th} data sample, j^{th} point. The network outputs a coordinates tuple for all the points of the object as $\{(\hat{x}^{(1)},\hat{y}^{(1)},\hat{z}^{(1)}),\dots,(\hat{x}^{(M)},\hat{y}^{(M)},\hat{z}^{(M)})\}$. To consider the ordering ambiguity of the network outputs in training, e.g., for $(x^{(1)},y^{(1)},z^{(1)})$, the network cannot determine which estimate gives lower MSE, between $(\hat{x}^{(1)},\hat{y}^{(1)},\hat{z}^{(1)})$ and $(\hat{x}^{(2)},\hat{y}^{(2)},\hat{z}^{(2)})$, we found proper orders by minimizing MSE over the permutation set P in (1) . With the help of minimization over P, the loss will be low as long as a trained network predicts the overall shape of the object, regardless of the order of the network estimates. In the training, we scaled down the true z coordinate values by 33.3 so that it is in the same range as coordinates x and y. This avoids the loss (1) from being dominated by z component of MSE loss, i.e., avoids training from being biased to z-coordinate estimation. We trained the network using Adam optimizer with the learning rate of 10^{-3}, the training batch size of 100, and 2000 epochs.

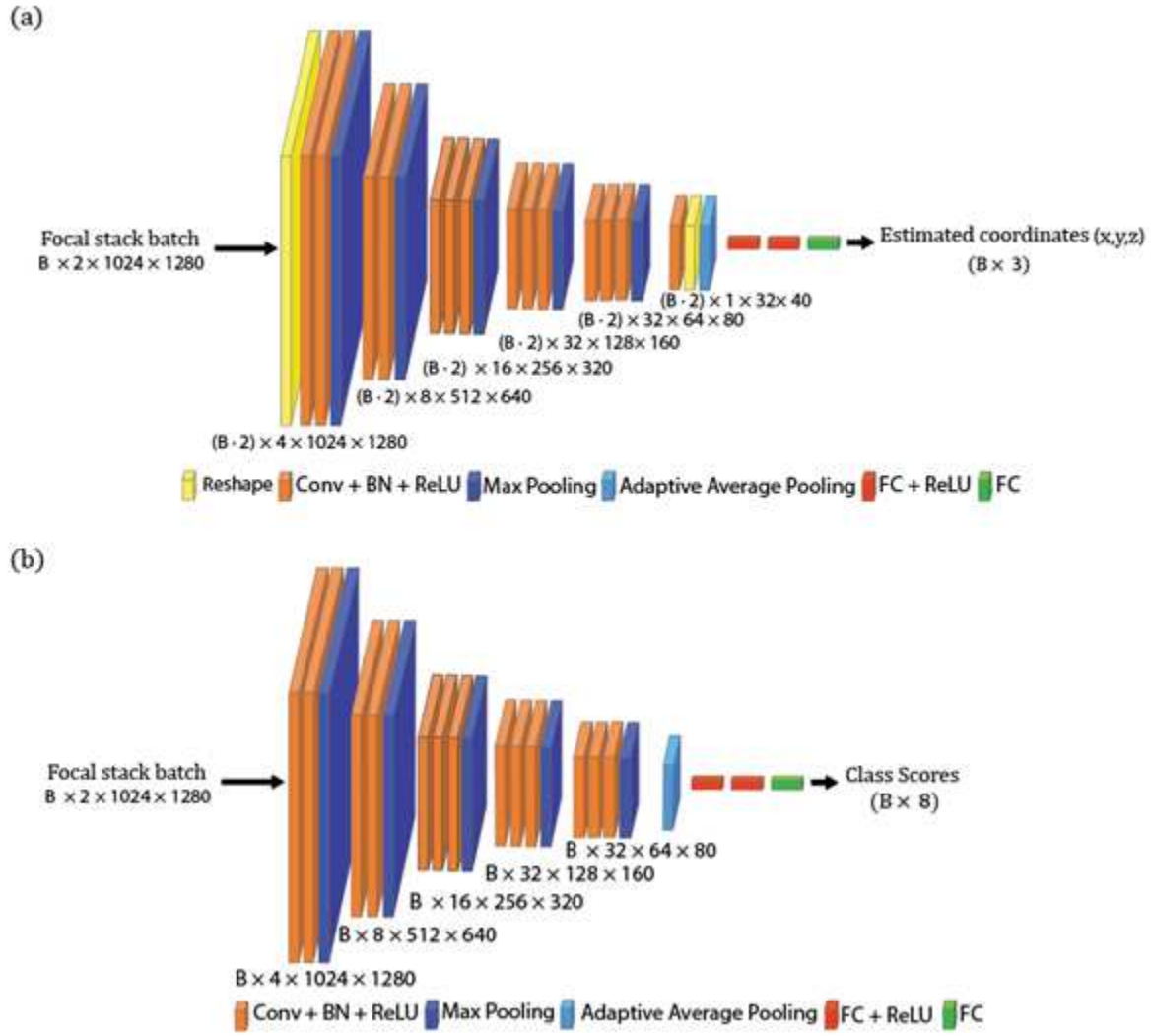
For tracking the two-point rotating object, we also trained the network by (1) and scaled the z coordinate values by 33.3. For training the network, we used Adam optimizer with the learning rate of 10^{-3}, the training batch size of 100, and 2000 epochs.

For extended object tracking and orientation estimation, we use two convolutional neural networks (CNNs) (Supplementary Fig. 6) similar to VGG-16[7]. The CNN shown in Supplementary Fig. 6(a) is used for the tracking. For each focal image, we first extract high-level feature maps with multiple convolution-batch normalization (BN)-ReLU-pooling layers. Then we apply the following procedure to extracted feature maps from all focal images: 1) concatenation of all feature maps along channel dimension, 2) average pooling, 3) flattening, and 4) feeding the output into fully connected layers (FC) that lead to final coordinates. The network is trained by minimizing the following MSE loss:

$$\frac{1}{3N}\sum_{i=1}^{N}\left(\hat{x}_i-x_i\right)^2+\left(\hat{y}_i-y_i\right)^2+\left(\hat{z}_i-z_i\right)^2,$$

where N is the number of training samples. In the training, we scaled the true z coordinate values to have the same range as x and y coordinates, for the same reason as in the multi-point object tracking. The CNN shown in Supplementary Fig. 6(b) is used for object orientation estimation. We consider the problem as a multi-class classification problem: the CNN takes focal stack as input and output scores that are used to classify the object orientations with eight different orientations. The network is trained by minimizing the cross-entropy loss.

We trained both CNNs with 13,188 training samples, using Adam optimizer with the initial learning rate of 10^{-4} (with learning rate decay by 0.3 at epoch 3, 5, 10, 20), the training batch size of 20, and 60 epochs and tested trained CNNs with 2,300 test samples. Due to the nondeterministic behavior of PyTorch[8], the training/testing is repeated three times for the orientation classification network and the best model is used. The orientation classification accuracies of the three runs are 95.35%, 96.61% and 99.35%.



Supplementary Figure 6. Convolutional neural network architectures for extended object tracking and orientation estimation. B is the general batch size of the data (e.g., in training, B is the training batch size; in testing with a single sample, $B = 1$). (a) Network for estimating extended object's spatial coordinates (x, y, z). (b) Network for estimating extended object's orientation.

V. Ranging Performance Comparison

We studied the effect of the detector resolution and spatial smoothing on the single-point object 3D ranging performance. Supplementary Table 1 summarizes the results. The resolution of the CMOS focal stack is varied to see its effect on the ranging performance: it can be seen by comparing horizontally the root mean square error (RMSE) in the 2nd, 3rd and 4th columns or in the 5th and 6th columns that higher resolution focal stack gives lower loss. Besides, note that spatially averaged results have lower loss, compared to those without averaging. This is because the noise from interference fringes is suppressed after applying spatial averaging.

	4×4 Graphene	4×4 CMOS	9×9 CMOS	32×32 CMOS	4×4 (Avg. 20) CMOS	9×9 (Avg. 20) CMOS
RMSE x	0.012	0.031	0.020	0.021	0.014	0.009
RMSE y	0.014	0.028	0.017	0.012	0.012	0.010
RMSE z	1.196	1.304	1.192	0.480	0.616	0.458

Supplementary Table 1. Single-point object 3D ranging RMSE (unit: mm) table on testing set. Avg. 20 means spatial averaging with window size 20 is performed on the raw high-resolution focal stack.

	4×4 Graphene	4×4 CMOS	9×9 CMOS	32×32 CMOS	4×4 (Avg. 20) CMOS	9×9 (Avg. 20) CMOS
2p2s	0.017	0.036	0.025	0.013	0.020	0.013
2p3s	0.019	0.033	0.022	0.013	0.019	0.012
3p2s	0.019	0.042	0.027	0.025	0.021	0.016
3p3s	0.021	0.041	0.029	0.028	0.022	0.017

Supplementary Table 2. Multi-point object 3D ranging RMSE (unit: mm) table of x on testing set. Avg. 20 means spatial averaging with window size 20 is performed on the raw high-resolution focal stack. First column encodes different object configurations, e.g., 2p3s means 2-point object with 3 possible shapes.

	4×4 Graphene	4×4 CMOS	9×9 CMOS	32×32 CMOS	4×4 (Avg. 20) CMOS	9×9 (Avg. 20) CMOS
2p2s	0.022	0.045	0.033	0.019	0.026	0.017
2p3s	0.025	0.039	0.028	0.018	0.025	0.015
3p2s	0.010	0.019	0.013	0.016	0.011	0.007
3p3s	0.019	0.035	0.026	0.027	0.021	0.016

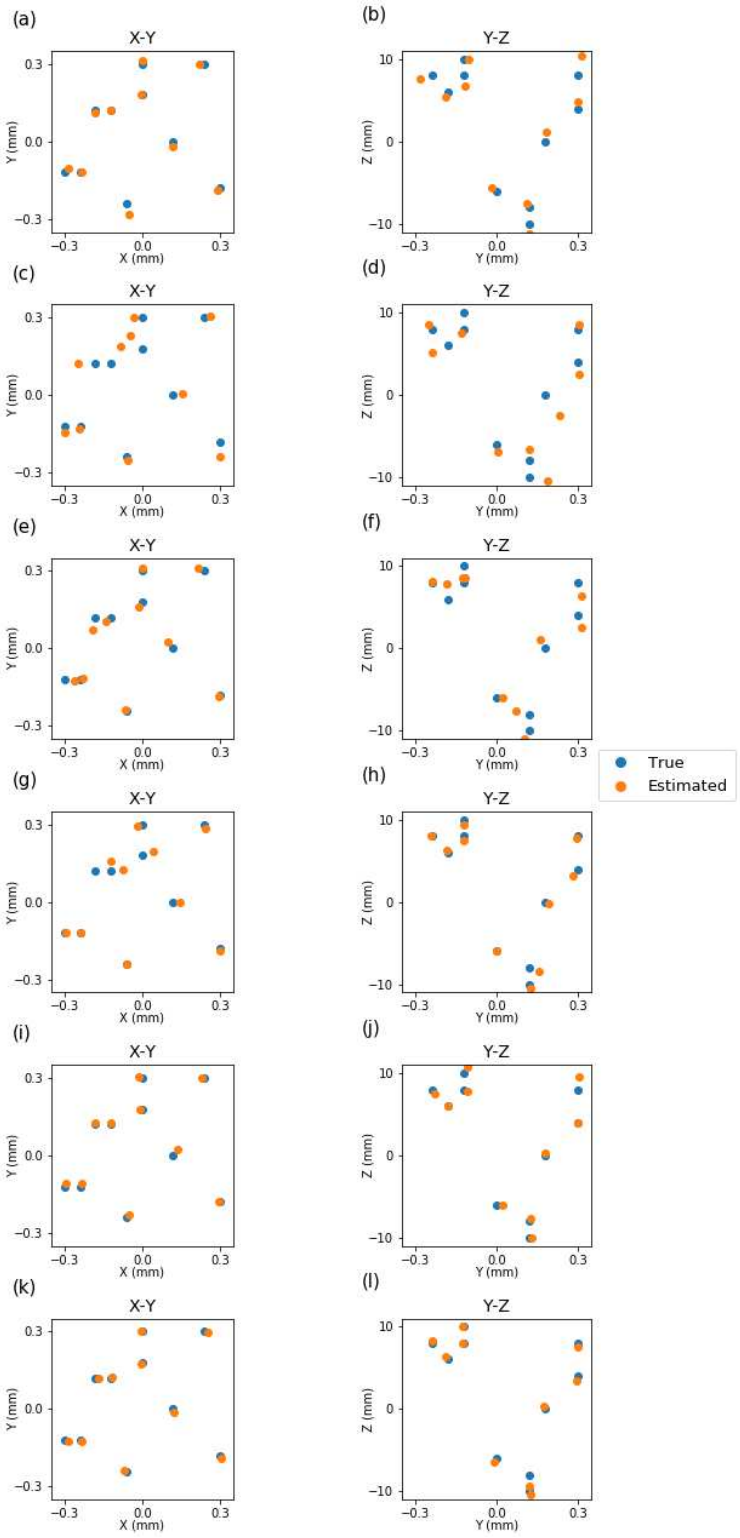
Supplementary Table 3. Multi-point object 3D ranging RMSE (unit: mm) table of y on testing set. Avg. 20 means spatial averaging with window size 20 is performed on the raw high-resolution focal stack. First column encodes different object configurations, e.g., 2p3s means 2-point object with 3 possible shapes.

	4×4 Graphene	4×4 CMOS	9×9 CMOS	32×32 CMOS	4×4 (Avg. 20) CMOS	9×9 (Avg. 20) CMOS
2p2s	0.685	1.073	0.759	0.349	0.557	0.371
2p3s	1.164	1.573	1.142	0.788	0.983	0.641
3p2s	0.793	1.328	0.876	0.715	0.750	0.470
3p3s	0.894	1.444	1.004	0.895	0.850	0.594

Supplementary Table 4. Multi-point object 3D ranging RMSE (unit: mm) table of z on a testing set. Avg. 20 means spatial averaging with window size 20 is performed on the raw high-resolution focal stack. First column encodes different object configurations, e.g., 2p3s means 2-point object with 3 possible shapes.

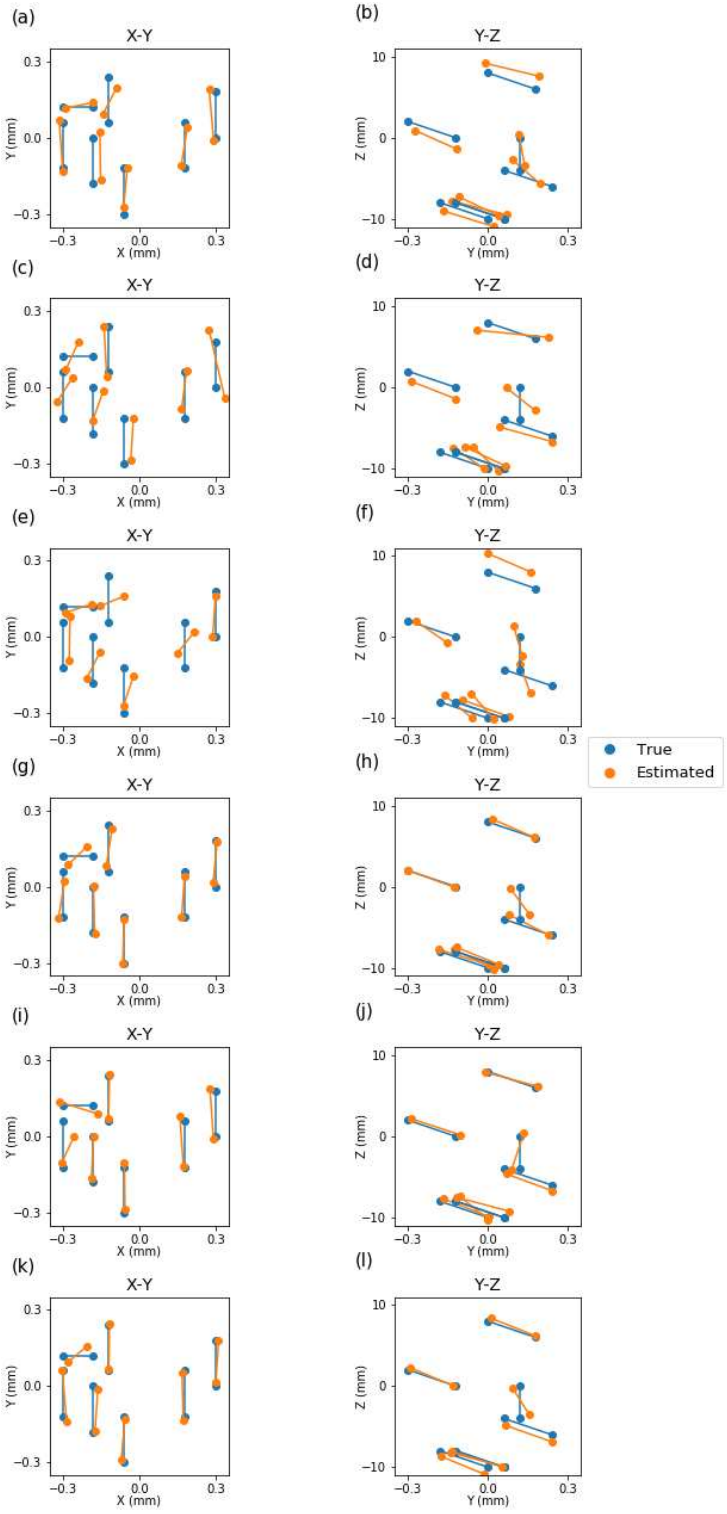
Supplementary Tables 2, 3, 4 summarize the study of the effect of the detector resolution and spatial smoothing on the multi-point object 3D ranging performance. Similar to the single-point object case, more pixels are useful in reducing the ranging error, as can be seen by comparing horizontally the RMSE in 2nd, 3rd and 4th columns or in the 5th and 6th columns. The spatial averaging is again helpful, as in the single object case, in reducing the estimation error.

The numerical results summarized in Supplementary Table 1-4 above are also illustrated graphically in Supplementary Fig. 7-11 below.

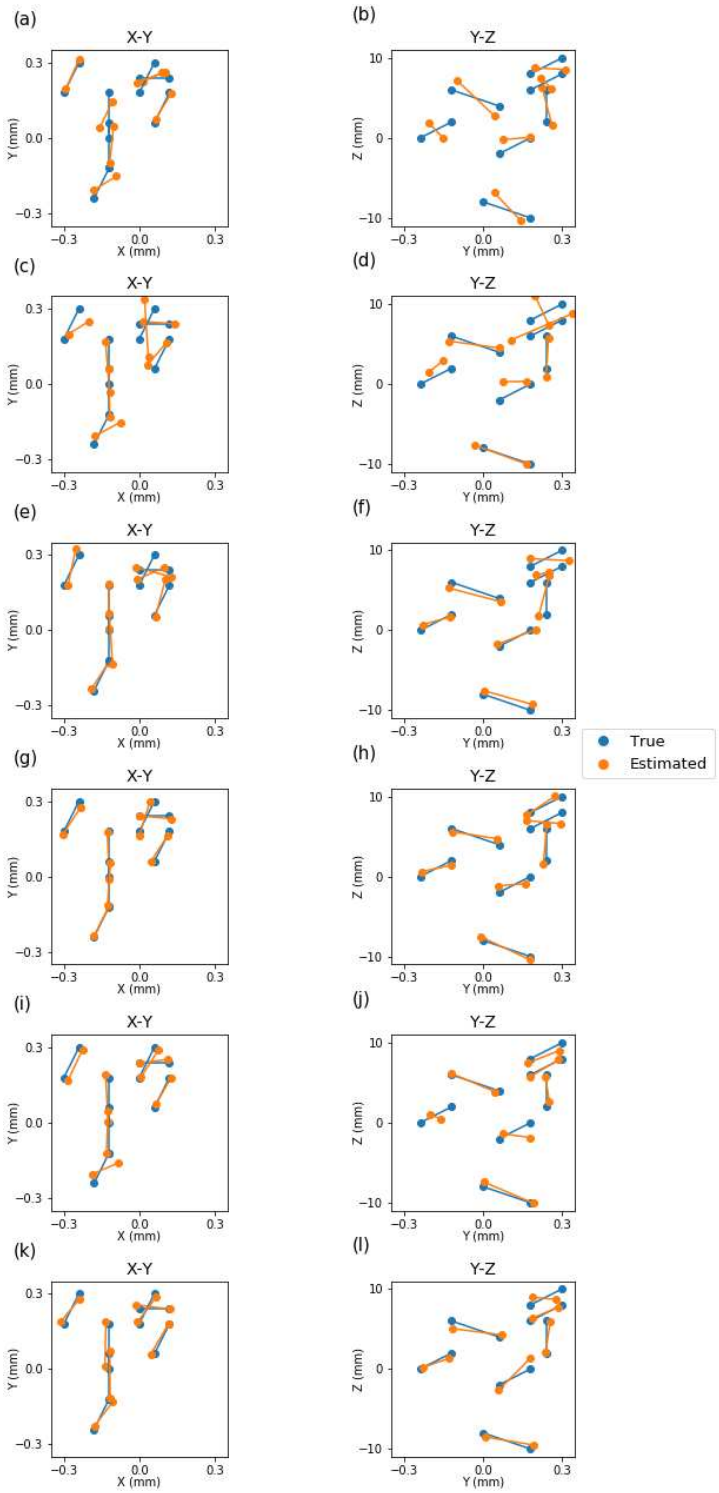


Supplementary Figure 7. Single-point object tracking performance (only 10 test samples are shown). Focal stack data from: (a-b) 4×4 transparent graphene detector. (c-d) 4×4 CMOS sensor. (e-f) 9×9

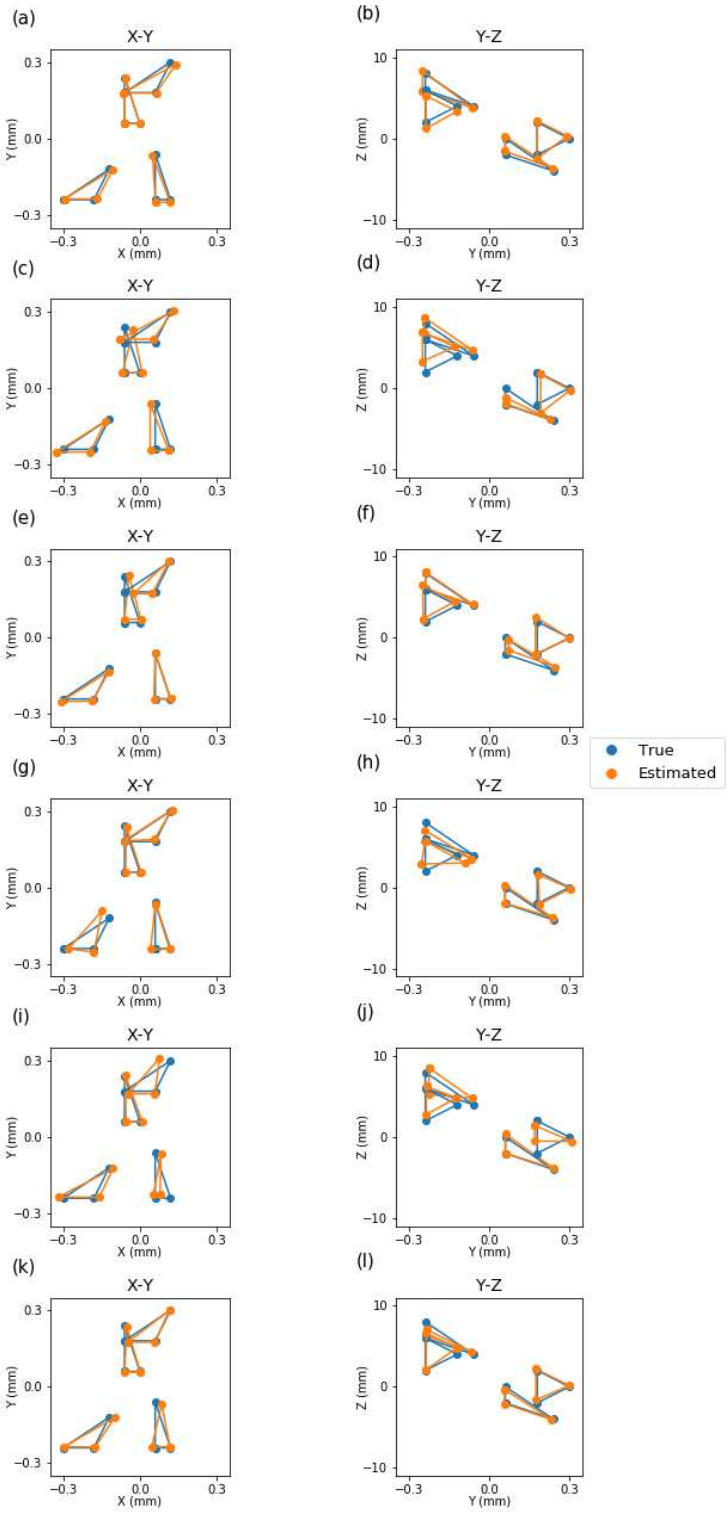
CMOS sensor. (g-h) 32×32 CMOS sensor. (i-j) 4×4 Avg. 20 CMOS sensor. (k-l) 9×9 Avg. 20 CMOS sensor.



Supplementary Figure 8. 2-point object with 2 possible shapes tracking performance (only 7 test samples are shown). Focal stack data from: (a-b) 4×4 transparent graphene detector. (c-d) 4×4 CMOS sensor. (e-f) 9×9 CMOS sensor. (g-h) 32×32 CMOS sensor. (i-j) 4×4 Avg. 20 CMOS sensor. (k-l) 9×9 Avg. 20 CMOS sensor.

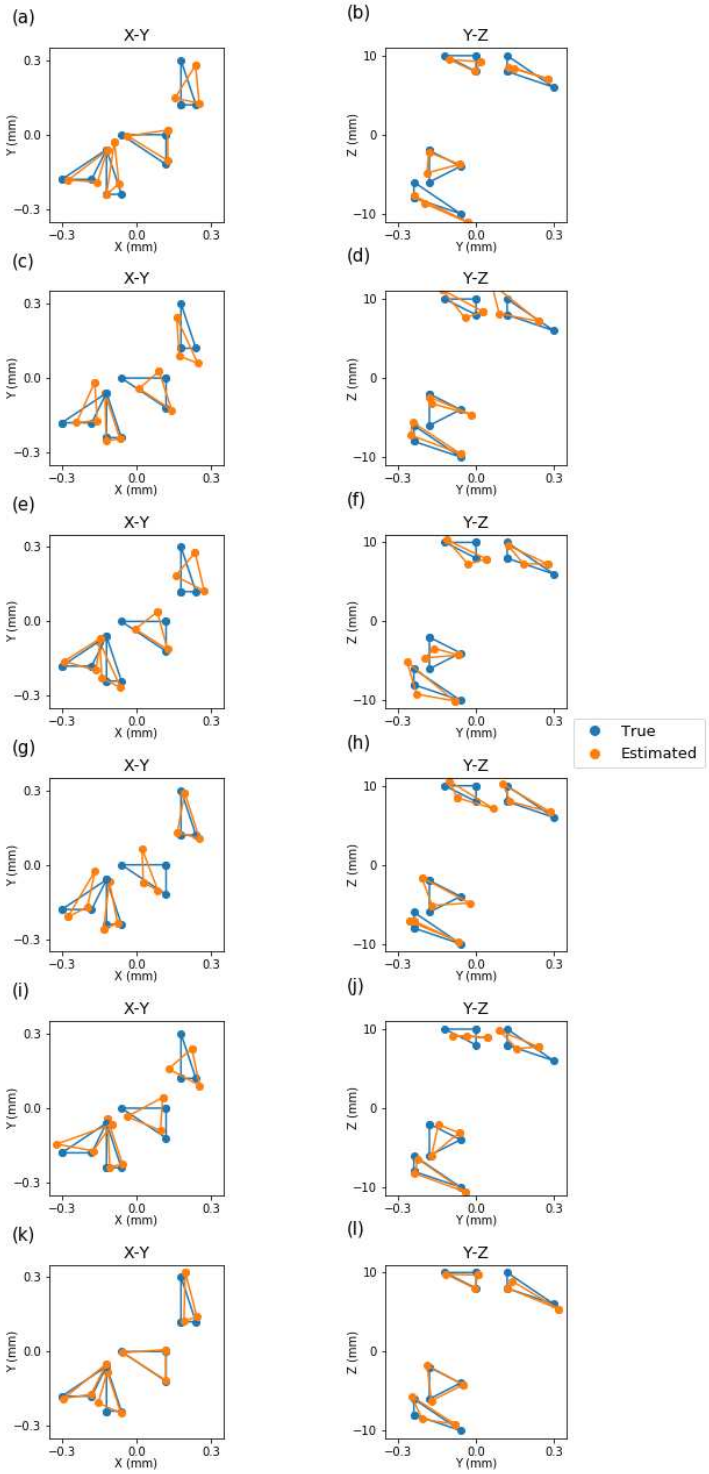


Supplementary Figure 9. 2-point object with 3 possible shapes tracking performance (only 7 test samples are shown). Focal stack data from: (a-b) 4×4 transparent graphene detector. (c-d) 4×4 CMOS sensor. (e-f) 9×9 CMOS sensor. (g-h) 32×32 CMOS sensor. (i-j) 4×4 Avg. 20 CMOS sensor. (k-l) 9×9 Avg. 20 CMOS sensor.



Supplementary Figure 10. 3-point object with 2 possible shapes tracking performance (only 4 test samples are shown). Focal stack data from: (a-b) 4 × 4 transparent graphene detector. (c-d) 4 × 4 CMOS sensor.

(e-f) 9 × 9 CMOS sensor. (g-h) 32 × 32 CMOS sensor. (i-j) 4 × 4 Avg. 20 CMOS sensor. (k-l) 9 × 9 Avg. 20 CMOS sensor.



Supplementary Figure 11. 3-point object with 3 possible shapes tracking performance (only 4 test samples are shown). Focal stack data from: (a-b) 4 × 4 transparent graphene detector. (c-d) 4 × 4 CMOS sensor.

(e-f) 9 × 9 CMOS sensor. (g-h) 32 × 32 CMOS sensor. (i-j) 4 × 4 Avg. 20 CMOS sensor. (k-l) 9 × 9 Avg. 20 CMOS sensor.

VI. Inference Time

Supplementary Table 4 shows the network inference time for point object tracking. The inference times for extended object position tracking and orientation estimation are 9.538 ms and 6.001 ms, respectively. We measured the inference time using a 1531 MHz Nvidia GeForce 1080 Ti GPU with 11G RAM.

	4 × 4	9 × 9	32 × 32
Single-point object	0.503 ms	0.512 ms	0.539 ms
2-point object	0.187 ms	0.190 ms	0.192 ms
3-point object	0.190 ms	0.189 ms	0.189 ms

Supplementary Table 4. Point object tracking inference time for different input focal stack resolutions.

Supplementary References:

[1] Lee, Seunghyun, et al. "Homogeneous bilayer graphene film based flexible transparent conductor." Nanoscale 4.2 (2012): 639-644.

[2] Zhang D, Cheng G, Xu Z, et al. Electrically tunable photoresponse in a graphene heterostructure photodetector[C]//2017 Conference on Lasers and Electro-Optics (CLEO). IEEE, 2017: 1-2.

[3] Balandin A A. Low-frequency 1/f noise in graphene devices[J]. Nature nanotechnology, 2013, 8(8): 549-555.

[4] Spietz L, Lehnert K W, Siddiqi I, et al. Primary electronic thermometry using the shot noise of a tunnel junction[J]. Science, 2003, 300(5627): 1929-1932.

[5] F. Rosenblatt, Psychol. Rev. 65, 386 (1958).

[6] Kingma, Diederik P., and Jimmy Ba. "Adam (2014), A method for stochastic optimization." Proceedings of the 3rd International Conference on Learning Representations (ICLR), arXiv preprint arXiv. Vol. 1412.

[7] S. Karen and A. Zisserman, ArXiv:1409.1556v6 (2014).

[8] https://pytorch.org/docs/stable/notes/randomness.html