

Supporting Information for “Machine-learned Light-field Camera Reads Facial Expression from High Contrast and Illumination Invariant 3D Facial Images”

Sang-In Bae¹, Sangyeon Lee¹, Jae-Myeong Kwon¹, Hyun-Kyung Kim¹, Kyungwon Jang¹, Doheon Lee¹, and Ki-Hun Jeong¹

¹Affiliation not available

October 21, 2021

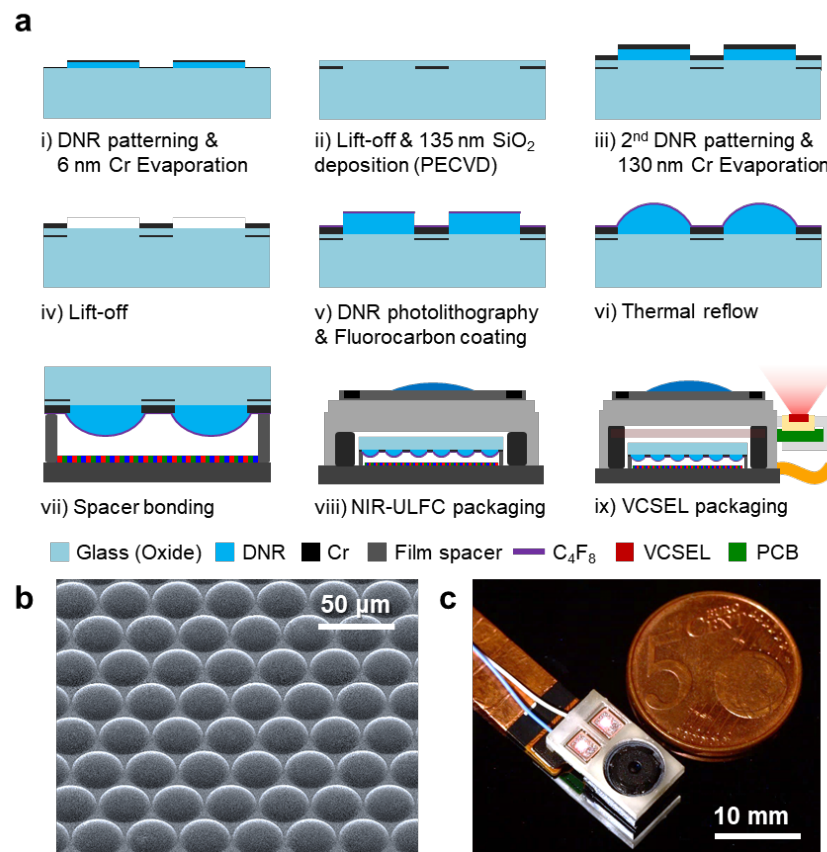


Figure S1. Detailed explanation of microfabrication step of fully integrated NIR-LFC. a) The wafer-level microfabrication of iMLA-AFF involves a thin Cr lift-off, and plasma enhanced chemical vapor deposition (PECVD) of SiO₂, and a thick Cr lift-off (6 nm Cr – 135 nm SiO₂ – 130 nm Cr), photolithographic patterning of DNR photoresist (DNR L300-D1, Dong-jin Semichem, Co., Ltd, Korea), and thermal reflow.

Note that a DNR photoresist exhibits both UV curable (Negative photoresist) and thermoplastic characteristic, suitable for metal lift-off as well as microlens formation. The hydrophobic coating of fluorocarbon (C_4F_8) effectively prevents the lateral expansion of microlenses on a metal surface during thermal reflow. iMLA-AFF are inversely bonded to an image sensor with a $60\ \mu\text{m}$ gap spacer and packaged to a compact objective lens by using a UV curable adhesive. The NIR-LFC is fully assembled by combining a $8.5\ \text{mm} \times 4.7\ \text{mm}$ printed circuit board with two VCSEL sources and VCSEL housing. b) A scanning electron microscope (SEM) of hexagonally arranged iMLA-AFF with $30\ \mu\text{m}$ in microlens diameter and $3\ \mu\text{m}$ in microlens gap. c) A photograph of fully packaged NIR-LFC. The camera module is connected to flexible extension cable and delivers raw image to Raspberry Pi 4(B). The total physical dimension of camera module is $8.5\ \text{mm} \times 14.0\ \text{mm} \times 5.6\ \text{mm}$.

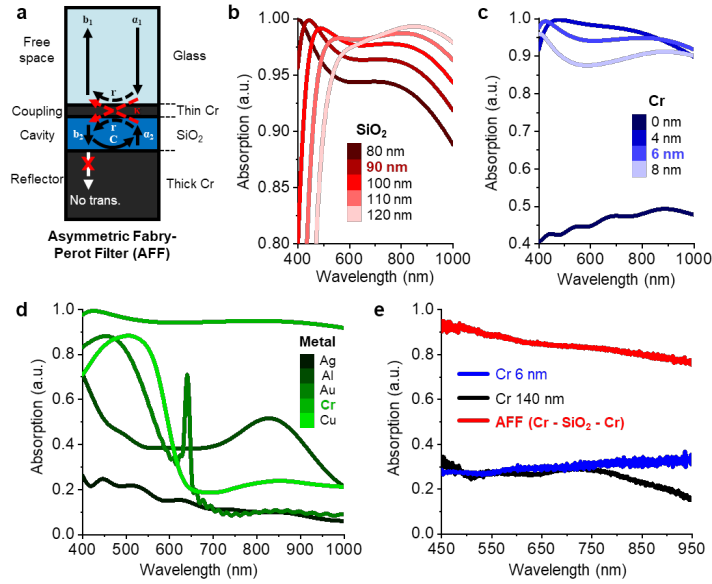


Figure S2. Absorption spectra of AFF from visible (VIS) to NIR wavelength region (400 – 1000 nm) based on the FDTD method and measurement absorption. a) Incident light passes through glass substrate, top absorptive Cr film, intermediate insulator layer of PECVD SiO₂, bottom reflective Cr, and air, in order. b) Absorption spectra at different SiO₂ thicknesses from 80 nm to 120 nm at wavelength from 400 - 1000 nm region. A 90 nm in thickness of SiO₂ exhibit broadband absorption over 0.93 from visible (VIS) to NIR wavelength region. c) The absorption of the upper Cr film is sensitive depending on the thickness compared to that of the lower film as a reflective metal due to its lossy nature. The peak absorption wavelength is blue-shifted as the bottom Cr layer thickness increases. d) Absorption spectra depending on a type of metal at VIS - NIR region. A Cr film has optically lossy and less dispersive characteristics, which can serve as bottom absorptive or top reflective layer. The interface reflection phase shifts become significantly different from 0 and π when a refractive index of top thin Cr film has comparable imaginary component (k) to real component (n) between 3.0 and 3.5 in the whole VIS-NIR region. e) Measured absorption spectra of AFF, 6 nm thin Cr film, and 90 nm thick Cr film, respectively, by using a spectrometer (SpectraPro® 2300i, Princeton Instruments). The AFF absorbs incident light over 0.8 in whole VIS-NIR wavelength region, but Cr films mostly transmit or reflect the light, which exhibit less than 0.4 in the same wavelength region.

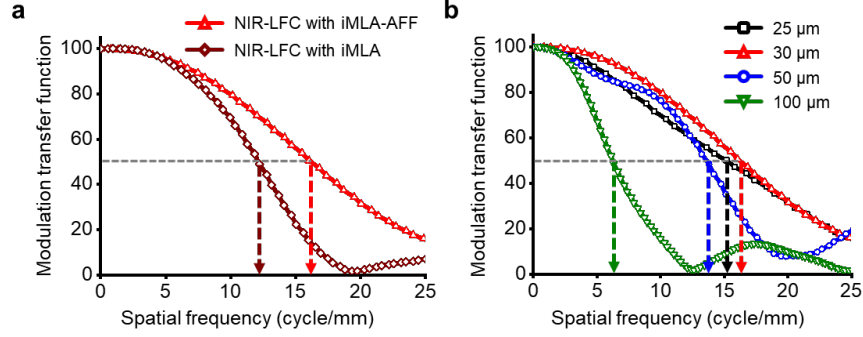


Figure S3. Modulation transfer functions of light-field renderings at different microlens conditions. A modulation transfer function (MTF) evaluates the fundamental spatial resolution performance and image sharpness of an imaging system from a black and white line pattern image. In particular, a MTF50, i.e., the 0.5 contrast value from the MTF curve, represents the resolving power of the images. The Quick MTF[®] software measures the quantitative MTF curves based on a slanted-edge method from the light-field rendering. a) MTF curves comparison of light-field renderings through NIR-LFC with iMLA-AFF (red triangles) and NIR-LFC with only iMLA (brown rhombus). The MTF50 values at each NIR-ULFCs demonstrate that the NIR-LFC with iMLA-AFF improves the image sharpness of light-field rendering by 34% ($MTF50_{iMLA-AFF} = 16.2$ cycle/mm, $MTF50_{iMLA} = 12.1$ cycle/mm). b) MTF curves of light-field renderings depending on microlens diameter (D_{MLA}) from 25 μm to 100 μm . The MTF50 value from 30 μm in D_{MLA} is 16.2 cycle/mm, which is higher than the MTF50 value from 25 μm in D_{MLA} , 15.1 cycle/mm. The smaller D_{MLA} results in more pixels in each sub-aperture image, which generally improves spatial resolution. But at D_{MLA} less than 30 μm , the spatial resolution decreases due to very low magnification. Therefore, the D_{MLA} of iMLA-AFF is determined to be 30 μm , which represents the highest MTF50 measurement in the CMOS image sensor of Sony IMX 219 (pixel size = 1.12 μm).

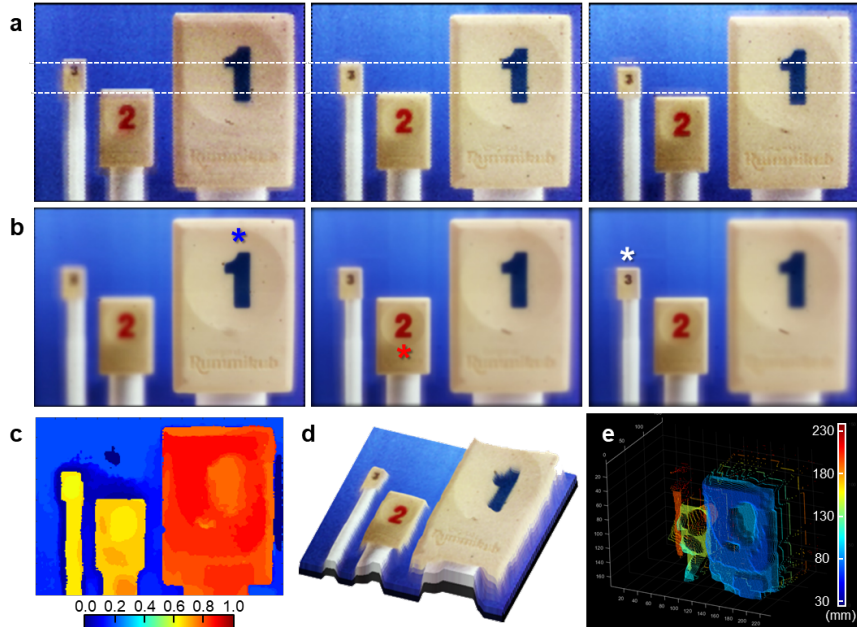
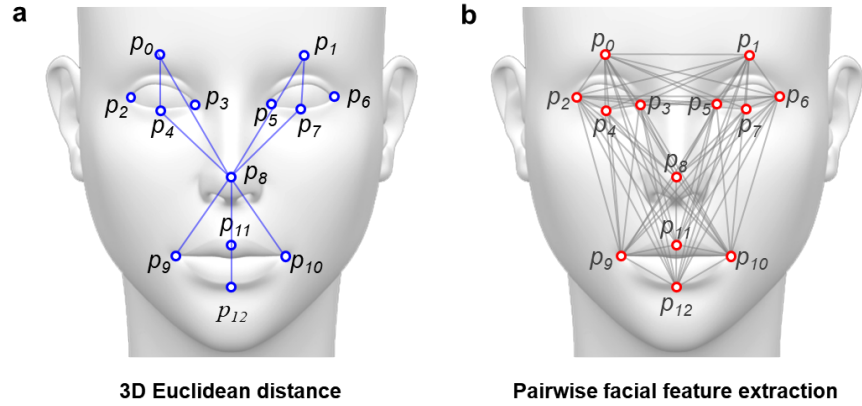


Figure S4. Computational light-field renderings and 3D depth reconstructions of numeric cube targets by using the fully integrated NIR-ULFC without NIR bandpass filter. Each of numeric

targets are located at 40 mm, 90 mm, and 180 mm from the NIR-LFC. a) Vertically shifted sub-aperture images with different viewing angles. The white dotted lines indicate the original position of numeric cube ‘3’ (center sub-aperture image), and the images on the left and right are obtained by changing the viewing angles up and down. b) Digital refocusing results of the targets, which focus on ‘1’ (left), ‘2’ (center), and ‘3’ (right), respectively. Asterisk symbols at each image indicate the focused point. c) Disparity map calculated by pixel-by-pixel cost volume filtering from the whole 27 x 27 sub-aperture images. d) 3D reconstruction of the targets after image warping by using a disparity map and a center sub-aperture image. The 2D image is warped on the disparity map. e) 3D depth map containing discrete depth information is calculated by using the disparity map obtained from c) and camera matrices, i.e., intrinsic parameter ($[K]$) and extrinsic parameter ($[R \ t]$) from geometric calibration steps. The calculated z-axis depth for each numeric target is 44 mm, 97 mm, and 191 mm, respectively, similar to those of actual positions.



	3D distance	3D distance
Facial Landmarks	$p_0, p_1, p_4, p_7, p_8, p_9, p_{10}, p_{12}$	$p_0 - p_{12}$
Feature	9 distances ($d_{p_0-p_4}, d_{p_1-p_7}, d_{p_4-p_8}, d_{p_7-p_8}, d_{p_0-p_8}, d_{p_1-p_8}, d_{p_8-p_9}, d_{p_8-p_{10}}, d_{p_8-p_{12}}$)	78 features ($d_{p_0-p_1}$ to $d_{p_{11}-p_{12}}$)
Purpose	Reconstruction error (RMSE) measurement	Multi-layer perceptron

Figure S5. Detailed description of feature landmark points. 13 different facial landmarks from 3D facial image are extracted by using the OpenFace[®] software, which extracts 68 different facial points and gaze direction from the face. a) Top eyebrow, lower eyelid, top nose, mouth end, and bottom lip of each right and left face are extracted, and 9 different Euclidean distances are calculated from both 3D depth map of the subject’s face. The 3D Euclidean distances measure the reconstruction errors of 3D depth maps acquired from different LFC depending on the various NIR filters. b) All 78 pairwise distance features of each subject’s different facial expression image are extracted ($d_{p_0-p_1}$ to $d_{p_{11}-p_{12}}$). The 78 features become a one input vector, and a total of 2,248 input vectors, including 281 input vectors from 4 different expressions repeated 3 times of 32 subjects and augmented random noises for robustness, are applied to the multi-layer perceptron (MLP) for reading facial expression.

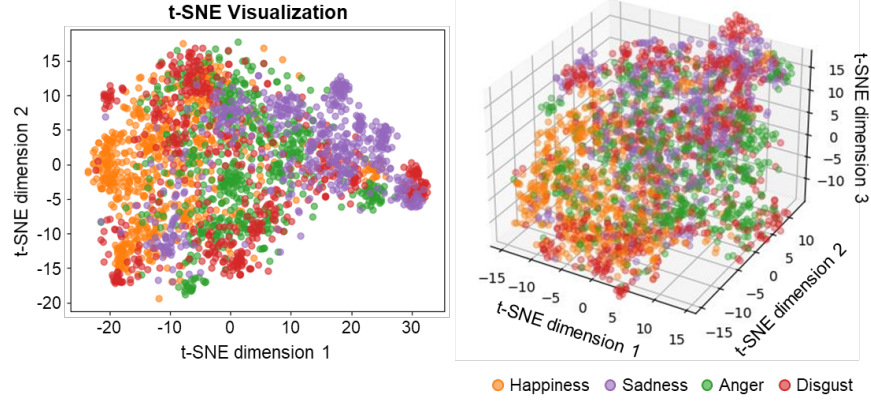


Figure S6. 2D and 3D Visualizations of facial data distribution. Pairwise distance vectors of 13 facial landmarks from facial images and augmented dataset are visualized in 2D plane using t-Stochastic Neighborhood Embedding (t-SNE) with perplexity 30. The clustered data points of each class imply that the facial expressions can be read by the pairwise distance vectors of the 3D facial images.

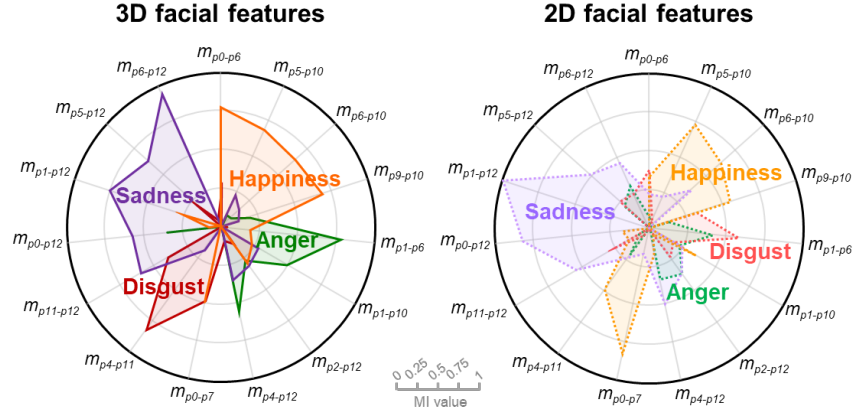


Figure S7. 2D and 3D radar distributions of expression-feature mutual information (MI). The MI quantitatively measures the mutual dependency between the distribution of 3D and 2D facial features (d_{pi-pj}) and the distribution of expression classes (0 or 1), classified through the MLP. The MI values from the 3D facial features exhibit a distinct and high mutual dependence for each facial expression, whereas those from the 2D features have an overlapped and low mutual dependence between other facial expressions except sadness. The four dominant mutual information values of both 3D and 2D facial features are normalized. More specific, the top 10 dominant features of each expression are as follows:

Happiness - d_{p0-p6} , d_{p9-p10} , d_{p6-p10} , d_{p5-p10} , d_{p7-p10} , d_{p0-p7} , d_{p2-p12} , d_{p2-p7} , d_{p9-p11} , d_{p0-p5}

Anger - d_{p4-p12} , d_{p1-p6} , d_{p2-p12} , d_{p1-p10} , d_{p7-p10} , d_{p8-p10} , d_{p2-p9} , d_{p0-p12} , d_{p4-p6} , d_{p1-p8}

Disgust - d_{p0-p7} , d_{p0-p6} , d_{p4-p11} , $d_{p11-p12}$, d_{p8-p9} , d_{p1-p4} , d_{p2-p12} , d_{p5-p12} , d_{p8-p11} , d_{p3-p7}

Sadness - d_{p6-p12} , d_{p1-p12} , d_{p5-p12} , d_{p0-p12} , d_{p2-p12} , d_{p7-p10} , $d_{p11-p12}$, $d_{p10-p12}$, d_{p2-p4} , d_{p0-p4}

Optical design of NIR-LFC

NIR-LFC has been fully packaged by placing an objective lens in a certain position (B_L , image sensor to objective lens distance) that covers the target object distance from 60 mm to 250 mm. Since the MLA project the objective lens image in LFC, an image-side depth-of-field (DOF) of an objective lens and an object-side DoF of MLA are matched at position a from the MLA plane. Since the MLA project the objective lens image in LFC, an image-side DOF of objective lens and an object-side DOF of MLA are matched at position a from MLA. The corresponding image-side depth-of-field DOF range of objective lens are denoted by a^+ and a^- planes. Therefore, the object-side DOF and its corresponding DOF range, i.e., the maximum and minimum object position (A_L^+ and A_L^-) with f_{OBJ} can be calculated using the object-side DOF range of MLA (a^+ and a^-).

A general concept to determine when a lens is in-focus or out-of-focus is to use the effective resolution of the imaging system. An effective resolution ratio (ERR) is defined by the ratio of effective resolution (R_e) and total resolution (R_t), and each of R_e and R_t are related to effective pixel size (s), minimal size of a projected point (s_0), and a single pixel size (p). D_I denotes total image sensor size.

$$R_e = \frac{D_I}{\max[|s|, s_0]}, R_t = \frac{D_I}{p},$$

$$ERR = \frac{R_e}{R_t} = \frac{p}{\max[|s|, s_0]}$$

An ERR of LFC (ERR_{LFC}) is determined by the image plane distance (B) and distance from image plane at a given p , microlens diameter (D_{MLA}), and microlens focal length (f_{MLA}),

$$ERR_{LFC} = \frac{1}{|v|} \bullet \frac{p}{\max\left[\left|D_{MLA} \left(\frac{B}{f_{MLA}} - \frac{1}{v} - 1\right)\right|, s_0\right]}, \quad (\text{for } |v| \geq 1)$$

A virtual depth, i.e., ratio of a and B , ($v = |a|/B$) is inversely proportional to ERR_{LFC} . From the above equations, the object-side DOF range of MLA (a^+ and a^-) are determined as,

$$a^- = \left[\frac{1}{f_{MLA}} - \frac{1}{B} \left(1 - \frac{s_0}{D_{MLA}}\right)\right]^{-1}, \quad a^+ = \left[\frac{1}{f_{MLA}} - \frac{1}{B} \left(1 + \frac{s_0}{D_{MLA}}\right)\right]^{-1}.$$

The image-side DOF range of LFC (DOF_{IS-LFC}) and a^{++} are calculated from ERR_{LFC}^- , i.e., an ERR value at a^- position.

$$ERR_{LFC}^- = \left|\frac{B}{a^-}\right| \bullet \frac{p}{\max\left[\left|D_{MLA} \left(\frac{B}{f_{MLA}} - \frac{B}{a^-} - 1\right)\right|, s_0\right]},$$

$$DOF_{IS-LFC} = \frac{1}{ERR_{LFC}^-} \bullet \frac{2pN}{\left|\frac{B}{f_{MLA}} - 1\right|} = a^{++} - a^-,$$

$$a^{++} = a^- + DOF_{IS-LFC}$$

Finally, minimum and maximum object distance, i.e., A_L^- and A_L^+ , are determined by using the thin lens equation of calculated a^- , a^{++} , f_{OBL} , and an objective lens position from IP (B_L). The DOF range of

LFC (DOF_{LFC}), i.e., object-side DOF range of objective lens, is the difference between maximum objective distance and minimum objective distance.

$$A_L^- = \left[\frac{1}{f_{\text{OBJ}}} - \frac{1}{(B_L - (a^- - B))} \right]^{-1}, A_L^+ = \left[\frac{1}{f_{\text{OBJ}}} - \frac{1}{(B_L - (a^{++} - B))} \right]^{-1}$$

$$\text{DoF}_{\text{LFC}} = A_L^+ - A_L^-$$

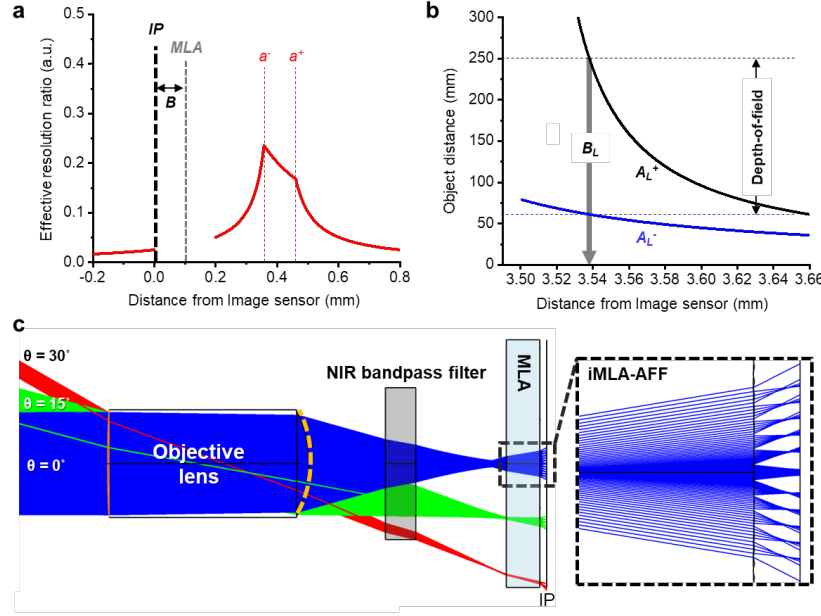


Figure S8. Optical design and ray diagram of NIR-LFC. NIR-LFC has been fully packaged by placing an objective lens in a certain position (B_L , image sensor to objective lens distance) that covers the target object distance from 60 mm to 250 mm. Since the MLA project the objective lens image in LFC, an image-side depth-of-field of an objective lens and an object-side DoF of MLA are matched at position a from the MLA plane. a) The corresponding image-side depth-of-field range of objective lens are denoted by a^+ and a^- planes (red dotted lines). Each of a^+ and a^- are determined by a thin lens equation of MLA focal length (f_{MLA}) and image plane distance (B), as

$$a^- = \left[\frac{1}{f_{\text{MLA}}} - \frac{1}{B} \left(1 - \frac{s_0}{D_{\text{MLA}}} \right) \right]^{-1}, a^+ = \left[\frac{1}{f_{\text{MLA}}} - \frac{1}{B} \left(1 + \frac{s_0}{D_{\text{MLA}}} \right) \right]^{-1}.$$

Note that $s_0 = \min[s_\lambda, p]$ is the minimal size of a projected point that can be resolved with the image sensor. s_λ and p denote the Airy disk and the pixel size, respectively. The optical design of NIR-LFC adopts a Keplerian imaging scheme, whose the image side DoF is generated in front of the image plane (IP) as positive signs ($a > 0$). b) The object-side DoF and its corresponding DoF range, i.e., the maximum and minimum object position (A_L^+ , black line and A_L^- , blue line) can be calculated using the thin lens equation of the object-side DoF range of MLA (a^+ and a^-) and f_{OBJ} .

$$A_L^- = \left[\frac{1}{f_{\text{OBJ}}} - \frac{1}{(B_L - (a^- - B))} \right]^{-1}, A_L^+ = \left[\frac{1}{f_{\text{OBJ}}} - \frac{1}{(B_L - (a^+ - B))} \right]^{-1}$$

$$\text{DoF}_{\text{LFC}} = A_L^+ - A_L^-$$

The B_L of NIR-LFC is determined to 3.539 mm that satisfying the target object distance. Note that the f_{OBJ} is 3.04 mm, B is 100 μm , and f_{MLA} and D_{MLA} are 75 μm and 30 μm , respectively. c) Ray diagram of NIR-LFC at different incident angles, 0° , 15° , and 30° , from 200 mm target simulated by the ZEMAX[®]. The objective lens is located 2.939 mm from the MLA, which considers MLA thickness of 500 μm and B of 100 μm .

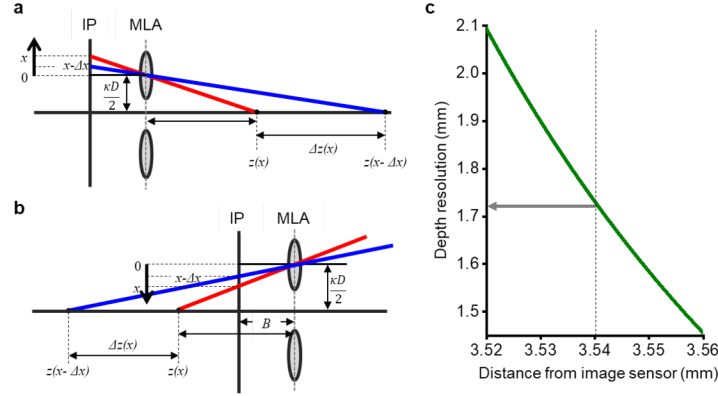


Figure S9. Depth resolution of NIR-LFC. As the position and orientation of microlens arrays (MLAs) is fixed at B , a simple triangulation and thin lens equation from MLAs and objective lens can estimate the depth resolution of LFC without the need of a 3D calibration of the system. We considered the depth resolution along a line parallel to the optical axis and halfway between the two microlenses, apart from κD . Note that a κ is a measure of the distance between the microlenses centers that are used for triangulation. An image point at distance x from the micro image center intersects the central bisecting line at point $z(x)$ is defined as $\frac{1}{x} \bullet \frac{\kappa B \Delta}{2}$

english2. A simple thin lens equation of $z(x)$, B_L , B , and f_L calculate the depth resolution in each a) Keplerian configuration and b) Galilean configuration. c) Depth resolution of NIR-LFC is determined by the pre-determined B_L , B , and f_L from Supplementary text and Fig. S6, and calculated as 1.7 mm. The κ is set to 2 assuming that the nearest distance between microlenses generate the highest depth resolution.