VII. Supplemental Material

A. Wave-optics Forward Model

The discrete linear forward light-field imaging model [24] can be formulated as

$$\mathbf{f} = H\mathbf{g} \tag{13}$$

where the vector $\mathbf{f}$ represents the light-field captured at the sensor plane, the vector $\mathbf{g}$ is the discretized version of the volume monitored by the microscope, and H is a measurement matrix whose coefficients are largely determined by the system's impulse response function, a.k.a. point spread function (PSF) of the light-field microscope.

The PSF denotes the generated pattern when an ideal point source passes through an optical system. It characterizes the features of the system. Different from the PSF of a conventional optical microscope that is usually a airy pattern for 2D imaging or a double-cone for 3-D imaging with translation-invariance property [47], the PSF of a light-field microscope has a more complex pattern which is translation-variant [24] and carries considerable information about the 3D positions of a point source in the volume and the physical property of the medium. Specifically, the pattern behind the MLA changes depending on the 3D positions of the point source. Thus, the imaging processing cannot be modeled as a convolution of a scene with a corresponding PSF, as is commonly done in the case of conventional image formation models [48]. Instead, the wavefront recorded at the sensor plane is described using a more general linear superposition integral [24]

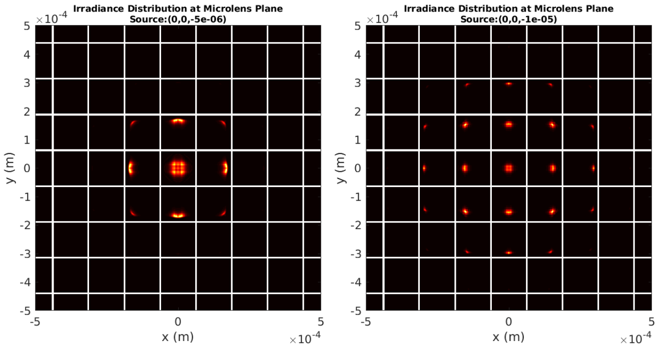
$$f(\mathbf{x}) = \int |h(\mathbf{x},\mathbf{p})|^2 g(\mathbf{p}) d\mathbf{p}, \tag{14}$$

where $\mathbf{p} \in \mathcal{R}^3$ is the position in a volume containing isotropic emitters whose combined intensities are distributed according to $g(\mathbf{p})$. When imaged, this volume gives rise to continuous 2D intensity pattern $f(\mathbf{x})$ at the image sensor plane. The optical impulse response $h(\mathbf{x},\mathbf{p})$ is a function of both the position $\mathbf{p}$ in the volume being imaged as well as the position $\mathbf{x} \in \mathcal{R}^2$ on the sensor plane. Some examples of such light-field PSF is shown in Fig. 17. PSF provides the basis for our localization and demixing algorithm.

Following [24], [47], a wave-optics forward model is developed to compute the PSF and illustrate light-field imaging process for a point source. Given a point source located at $\mathbf{p} = (p_1, p_2, p_3)$, the Debye theory is utilized to calculate the light-field at a point close to the point of convergence of a wave. Specifically, the analytical model for the wavefront at the native object plane generated by a point source at $\mathbf{p} = (p_1, p_2, p_3)$ can be written as:

$$U_o(\mathbf{x},\mathbf{p}) \propto \int_0^\alpha P(\theta) J_o\left(v\frac{\sin(\theta)}{\sin(\alpha)}\right) \exp\left(-ju\frac{\sin^2(\theta/2)}{2\sin^2(\alpha/2)}\right) \sin(\theta)\,\mathrm{d}\theta \tag{15}$$

where, the auxiliary variables v, u represent normalized radial and axial optical coordinates, defined as $v = k\sqrt{(x_1-p_1)^2+(x_2-p_2)^2}\sin\alpha$ and $u = 4knp_3\sin^2(\alpha/2)$, respectively, where n denotes the refractive index of the material, e.g. water or oil in which specimen is immersed. J_o denotes the zero-th order Bessel function of the first kind; $\alpha = \sin^{-1}(NA/n)$ denotes the half-angle of the numerical aperture NA; $k = 2\pi/\lambda$ denotes the angular wavenumber; finally, $P(\theta)$ denotes the apodization function of the microscope, e.g., $P(\theta) = \sqrt{\cos(\theta)}$ for Abbe-sine corrected objectives. Note that, this equation only holds for low to moderate NA objectives.



(a) PSF at depth of 5 μm.

(b) PSF at depth of 10 μm.

Figure 17. Simulated PSF for an ideal point source at different depths.

Given $U_o(\mathbf{x},\mathbf{p})$, the lightfield $U_i(\mathbf{x},\mathbf{p})$ at the native image plane of a 4-f system is formulated as the inverted and stretched version of $U_o(\mathbf{x},\mathbf{p})$:

$$U_i(\mathbf{x},\mathbf{p}) = U_o(-\mathbf{x}/M,\mathbf{p}) \tag{16}$$

where M denotes the magnification of the 4-f system.

Next, we model the wavefront passing through a MLA. The MLA used in our experiment contains square-truncated lenslets with squared aperture and a 100% fill factor. Considering a single lenslet centered on the optical axis with focal length f_{ML} and pitch d, its transmittance $\phi(\mathbf{x})$ is defined as

$$\phi(\mathbf{x}) = P(\mathbf{x}) \exp\left(\frac{-jk}{2f_{ML}}\|\mathbf{x}\|_2^2\right) \tag{17}$$

where, $P(\mathbf{x})$ denotes the pupil function, a.k.a amplitude mask, representing the lenslet aperture, e.g. $rect(\mathbf{x}/d)$ for a square lenslet, and the term with exp denotes the phase mask representing the refraction of light through the lenslet itself.

Accordingly the lens mask of a MLA $\Phi(\mathbf{x})$ can be described as a convolution of 2D Dirac impulses, a.k.a. 2D Dirac comb function $comb(\mathbf{x}/d)$, with $\phi(\mathbf{x})$:

$$\Phi(\mathbf{x}) = \phi(\mathbf{x}) \odot comb(\mathbf{x}/d)\,. \tag{18}$$

Finally, the light-field $h(\mathbf{x},\mathbf{p})$ at the sensor plane is obtained by multiplying $U_i(\mathbf{x},\mathbf{p})$ by the lens mask $\Phi(\mathbf{x})$ and then propagating the result from the MLA to the sensor plane using the paraxial approximation:

$$h(\mathbf{x},\mathbf{p}) = \mathcal{F}^{-1}\{\mathcal{F}\{U_i(\mathbf{x},\mathbf{p})\Phi(\mathbf{x})\}G(\hat{\mathbf{x}})\} \tag{19}$$

where, $G(\hat{\mathbf{x}}) = \exp(-\frac{i}{4\pi}\lambda f_{ML}\|\hat{\mathbf{x}}\|_2^2))$ denotes the transfer function for propagating U_i from the MLA plane to the sensor plane using a Fresnel diffraction integral when the Fresnel number of lenslets is between 1 and 10. $\hat{\mathbf{x}}$ are spatial frequencies with coordinates in the Fourier domain along the x_1 and x_2 directions in the sensor plane.

B. *Experimental Setup and Results*

Experimental Setup.

The designed LFM system (shown in Fig. 7) is modified from a fluorescence microscopy by inserting a MLA (pitch 125 µm, f/10, RPC Photonics) at the imaging plane of an objective lens (25×, $NA = 1.0$, Olympus) and tube lens (180nm, Thorlabs) with a CMOS sensor (ORCA Flash 4, Hamamatsu) placed at its back focal plane. By the principles of light-field imaging, each lenslet records the angular distribution of light rays, therefore such design allows to capture both position and direction of propagation of light rays with a single-shot in a 2D intensity image.

Specifically, since the resolving power of a multi-lens optical system is governed by the smallest Numerical Aperture (NA) among its lenses, we need to select the MLA with an appropriate F-number to ensure that its NA matches with the (imaging side) NA of the objective and thus it does not limit resolution of the microscope. For example, given a water immersion objective with magnification 25× and $NA = 1.0$, its imaging side NA is 0.04 and thus the F-number of the MLA should be smaller than 12.5[2]. We select a MLA with F-number f/10 to ensures that the micro images behind the lenslets tile the image plane without overlapping or leaving spaces in between them [23], [24].

As noted before, capturing 4D information of light rays using a 2D sensor leads to a tradeoff between spatial resolution (i.e. the number $N_k \times N_l$ of lenslets in the MLA) and angular resolution (i.e. the number $N_i \times N_j$ of pixels behind each lenslet) which is controlled by a few parameters. In particular, the total resolution $N_i \times N_j \times N_k \times N_l$ of the LFM is limited by the number of resolvable sample spots in the specimen [23]. Here, we adopt a commonly used metric Sparrow limit [49] to measure the upper limit of the resolution. Sparrow limit is defined as the smallest spacing between two points on the specimen such that the intensity along a line connecting their centers in the image barely shows a measurable dip. It is expressed as a distance on the intermediate image plane:

$$R_{obj} = \frac{0.47\lambda}{NA} M \tag{20}$$

where λ is the wavelength of light. Since the MLA does not change the number of resolvable spots, the angular and spatial resolution upper limit can be derived according to the relation:

$$\lceil N_i \times N_j \rceil = \frac{W \times H}{R_{obj}} \tag{21}$$

where $W \times H$ are the size of a lenslet and $\lceil N_i \times N_j \rceil$ are the number of resolvable spots behind each lenslet, representing the upper limit of the angular resolution.

Taking our system for example, under red light (650 nm), a $25\times/1.0NA$ objective has Sparrow limit $R_{obj} = 7.64$µm. The MLA used in our LFM is composed of lenslets with square aperture and the size of each lenslet $W \times H$ is equal to the lenslet pitch $d = 125$µm. Accordingly, the upper limit of the angular resolution is $\lceil N_i \times N_j \rceil = d/R_{obj} \times d/R_{obj} \approx 16.36 \times 16.36$. Our LFM system has a 4-f relay system with the focal length f_{obj} of the objective as 7.2 mm and the focal length f_{tl} of tube-lens as 180 mm, leading to a magnification factor $M = 25\times$. Accordingly, the pixel size at the sample plane is $d/M = 125$µm$/25 = 5$µm, which is the upper limit of the spatial resolution on the specimen. Furthermore, since the focal length f_{ML} of the MLA is 1250 µm, the sampling angular range for each lenslet is $\theta = 2\arctan(d/(2f_{ML})) \approx d/f_{ML} = 125$µm$/1250$µm $= 0.1 = 5.73°$, while the angular resolution in degree is $\delta\theta = \theta/N_i = 0.1/19.2 = 0.2986°$. So it can be noted that microlens-based light-field imaging performs dense sampling in the angular space.

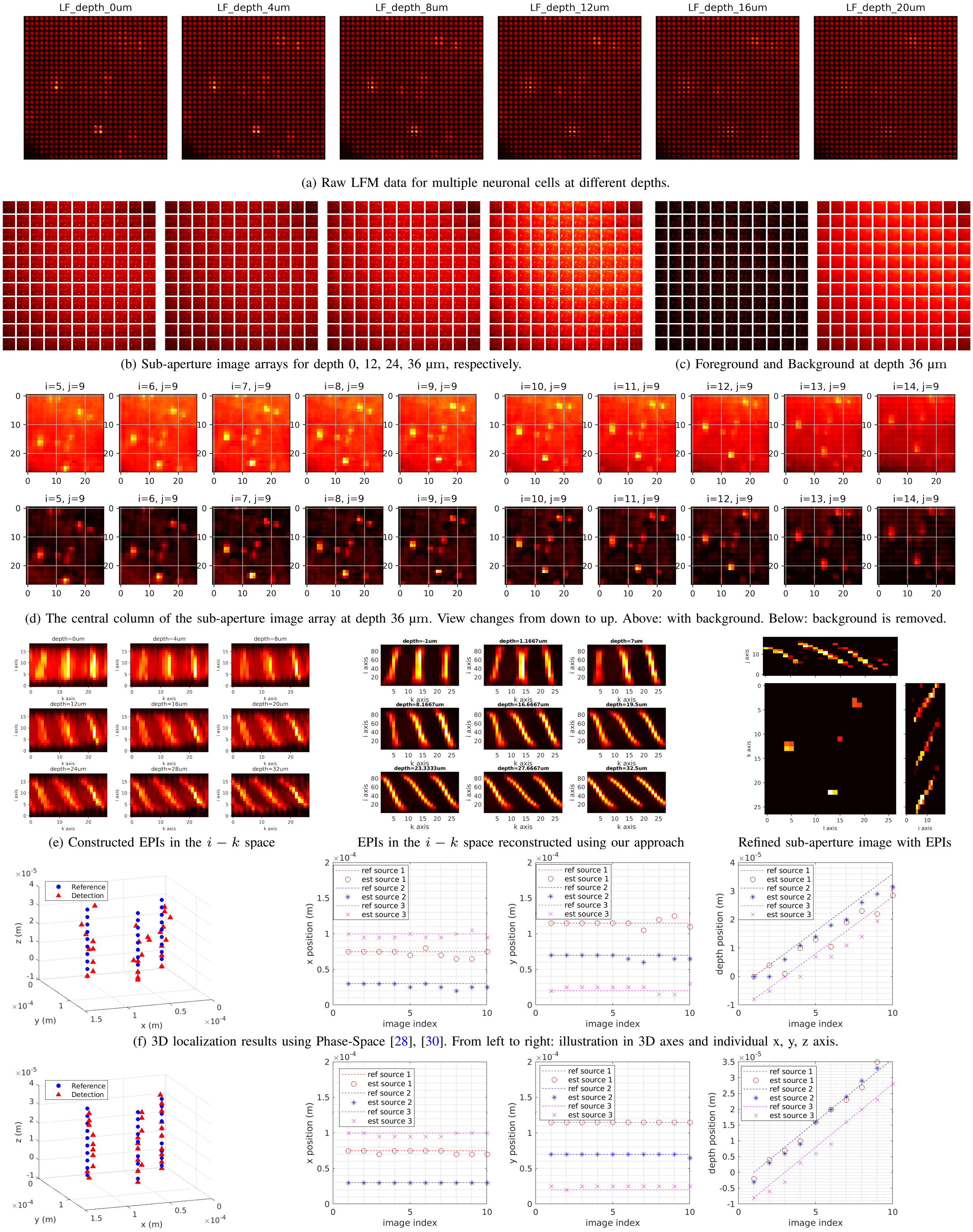
Once the resolution upper limit is known, we can select an appropriate sensor to satisfy the requirements. In our system, the size of a single pixel in our selected CMOS sensor (ORCA Flash 4, Hamamatsu) is 6.5 µm, leading to $N_i \times N_j = d/6.5 \times d/6.5 = 19.2 \times 19.2$ pixels behind each lenslet. This is higher than the upper limit, therefore the sensor satisfies the requirement and will not hamper the resolving power of the system. In addition, the CMOS sensor has 2048×2048 pixels, leading to a field of view (FOV) with 13.31mm × 13.31mm at the imaging side and equivalently 532.48µm × 532.48µm at the sample plane. Accordingly, the spatial size of a sub-aperture image is $N_k \times N_l = 106 \times 106$ pixels.

Experimental Results.

We conducted additional experiments for the scattering case involving multiple cells obtained from a genetically encoded mouse and located at different depths away from the focal plane, as shown in Fig. 18. Subfigure (a) and (b) show that each raw 2D light-field image is converted into the standard 4D format and then the pixels are re-arranged into a sub-aperture image array. Subfigure (d) shows that the view changing phenomenon can be observed in a sequence of sub-aperture images. The constructed and reconstructed EPIs are shown in subfigure (e). The rightmost subfigure reflects the relation between the EPIs and the corresponding sub-aperture image at the center view. Subfigure (f) and (g) compare the 3D localization performance of Phase-Space [28], [30] and our approach. The average RMSE using Phase-Space [28], [30] is 4.63 µm, 4.89 µm, 4.15 µm, for x, y and z positions, respectively. In contrast, the average RMSE using our approach is 2.23 µm, 2.11 µm, 1.69 µm for x, y and z positions, respectively. This set of experiments also verifies that our approach outperforms the state-of-art method Phase-Space [28], [30] with better 3D localization performance.

We also evaluated 3D localization performance after randomly changing the horizontal and vertical positions. As shown in Fig. 19 where the blue round points represent the reference and the red triangle points represent the detection results. For non-scattering case. The average RMSE of 3D localization is 2.06 µm, 2.12 µm and 2.39 µm for x, y and z positions, respectively. For scattering case. The average RMSE of 3D localization is 1.73 µm, 2.00 µm, 1.71 µm for x, y and z positions, respectively.

[2] Numerical aperture can be converted to F-number (f/stop) using the approximate formula F-number = $1/(2NA)$.



(a) Raw LFM data for multiple neuronal cells at different depths.

(b) Sub-aperture image arrays for depth 0, 12, 24, 36 µm, respectively.

(c) Foreground and Background at depth 36 µm

(d) The central column of the sub-aperture image array at depth 36 µm. View changes from down to up. Above: with background. Below: background is removed.

(e) Constructed EPIs in the $i-k$ space

EPIs in the $i-k$ space reconstructed using our approach

Refined sub-aperture image with EPIs

(f) 3D localization results using Phase-Space [28], [30]. From left to right: illustration in 3D axes and individual x, y, z axis.

(g) 3D localization results using our approach. From left to right: illustration in 3D axes and individual x, y, z axis.

Figure 18. Scattering case for multiple cells. (a) Raw LFM images of multiple neuronal cells (from a genetically encoded mouse) at different depths away from the focal plane. (b) Sub-aperture image arrays for different depths. (c) The separated foreground and background of a sub-aperture image array via SVD. (d) From a column of the sub-aperture image array, it is noticed that the positions of the bright area are shifting, which means the view direction is changing vertically. (e) Left and middle: constructed and reconstructed EPIs. Right: Relation between EPIs and the sub-aperture image at the center view. (f) 3D localization performance in terms of RMSE using Phase-Space [28], [30]. The average RMSE is 4.63 µm, 4.89 µm, 4.15 µm, for x, y and z positions, respectively. (g) 3D localization performance in terms of RMSE using our approach. The average RMSE is 2.23 µm, 2.11 µm, 1.69 µm for x, y and z positions, respectively. Best seen by zooming on a computer screen.

C. Additional Discussion

In this section, we provide additional discussion related to our approach.

1) Differences from phase-space based methods [28]–[30]: Epi-polar plane image is a way to reveal the structure of light-field in an appropriate space. EPI turns out to be similar to other concepts, such as phase space, spatial-angle space, etc. which have been used in some work [28]–[30] to leverage spatial and angular information simultaneously. However, we also need to point out that EPI or phase-space are, in fact, just two types of representations that can be used to manipulate multi-view images and that there are alternative ways to use these representations. We highlight that it is the proposed paradigm of combing specialized EPI dictionary with convolutional sparse coding that makes our localization approach distinctive.

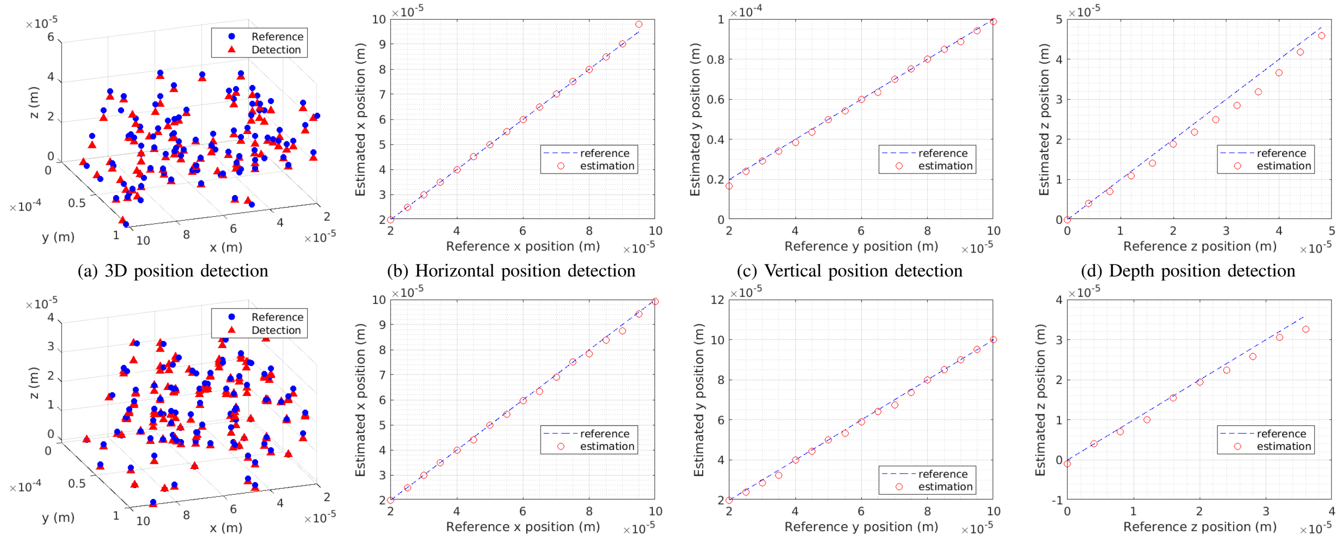
First and foremost, different from the phase-space based method [28], [30], we capitalized on the shift-invariance property of EPI via convolutional sparse coding, which considerably reduces the dictionary size, computational complexity, and improves the upper limit of localization resolution at depth. Specifically, the shift of a point source in a transverse dimension, e.g. along x or y coordinate, corresponds to the shift of the epipolar line along the same coordinate in the EPI. This shift-invariance property enables us to only consider the depth range when synthesizing dictionary elements, as transverse shift can be revealed by a convolution operation. It also accounts for why convolution is an effective operation to search for specific patterns and to perform pattern recognition. In contrast, without exploiting the shift-invariance property, the number of elements in a dictionary can increase dramatically along with increasing transverse dimensions. We note that [28], [30] did not exploit such shift-invariance property in their forward model. Instead, they considered the spatial distribution of sources and created a forward model for each point source in the 3D space to describe the light-field. Therefore, the dictionary size can be huge, and the computational complexity can be prohibitively high.

For example, covering a 3D space of $1000 \times 1000 \times 1000 \mu\text{m}^3$ with a resolution of $10\mu\text{m}$ will require to compute 10^6 forward models, i.e. 10^6 atoms in a dictionary, without exploiting the shift-invariance property. Therefore, the dictionary size is huge, and the computational complexity is prohibitively high. In contrast, if the shift-invariance property is used, only the depth dimension, that is 100 different depths in the example, need to be considered and only 100 atoms are required in a dictionary. So for this example a 4D phase-space dictionary may require 10000 times more memory than our dictionary.

Second, our dictionary elements are not point spread functions (PSF) as in [28], [30]. Each dictionary atom in our approach represents light-field from a ball-shaped volumetric source with a reasonable radius at a specific depth. Therefore, our dictionary elements are not point spread functions (PSF) for ideal point sources without radius. This is another significant difference from the model exploited in [28], [30]. Apparently, it is more practical and realistic to model ball-shaped volumetric sources instead of ideal point sources when the targets are neuron cells or fluorescent beads. A notable benefit of using a more realistic dictionary model is that it contributes to enhanced sparsity because fewer atoms are required for faithful representation. This, in turn, leads to better robustness and localization performance. As shown in Fig. 20, given the same ball-shaped sources and under the same sparsity regularization, convolutional sparse coding with respect to our dictionary gives sparser and more structured coefficients, as well as lower root mean squared error (RMSE) than using a dictionary consisting of PSFs.

Third, in addition to wave-optical effects, our light-field model also considers the effects of main lens and the microlens array of the microscopy system along the whole light-field propagation path, which ensures our model is more similar to the real observations than the model in [28], [30]. As mentioned before in Section V (Experiments), the main lens together with the relevant 4F system ensures that the electromagnetic field is approximately band-limited in space. That also results in non-uniform light distribution in the imaging plane so that the light density is the largest in the central region and becomes smaller for areas far away from the center. This accounts for why epipolar lines in our EPIs tend to be thicker in the central region and thinner at the two ends, as shown in Fig. 21 (b) (or Fig. 6 in the manuscript), in particular for out-of-focus sources, e.g. at a depth of 20 um. This phenomenon also matches real light-field observations, as shown in Figure 10 (d) and Figure 11 (e). In addition, the blurring and downsampling effects from the microlens array and associated pixels behind each lenslet are also incorporated into our model. In contrast, Phase-Space [28], [30] incorporates the wave-optical and geometric effects into their model using a phase-space Wigner function (and its Fourier spectrogram) so that the light propagation in space can be easily represented by a simple shearing operation in phase-space. [30] measures the Fourier spectrogram of Wigner function by scanning an aperture to capture the local power spectrum for each position, while collecting real-space images at the same time. However, the effects of the main lens and the microlens array were ignored. The fact that these effects were not fully incorporated may account for why their phase-space dictionary elements are straight lines with uniform shearing everywhere, as shown in Fig. 21 (a). It can be noticed that the simulated phase-space elements do not resemble real phase-space observations, in particular at deeper positions where the real observations exhibit an ’S’-shape due to distortion and aberrations from the lenses. These mismatches between the phase-space dictionary and real light-field observations may cause localization errors. In contrast, our EPI dictionary elements better resemble real observations. Such accuracy contributes to our better 3D localization performance, as it produces better convolutional sparse coding performance and algorithm robustness.

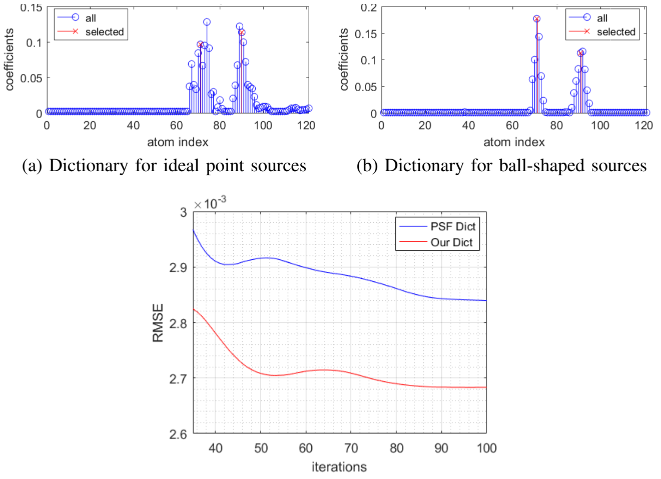
Fourth, the purpose of our dictionary is different from [28]. The dictionary in [28] consists of a set of footprints, each corresponding to a light-field image with only one cell in it, referred to as a light-field signature. The sparse coding with respect to the dictionary gives coefficients which correspond to the magnitude of the functional data, i.e. calcium transients.



(a) 3D position detection (b) Horizontal position detection (c) Vertical position detection (d) Depth position detection

(e) 3D position detection (f) Horizontal position detection (g) Vertical position detection (h) Depth position detection

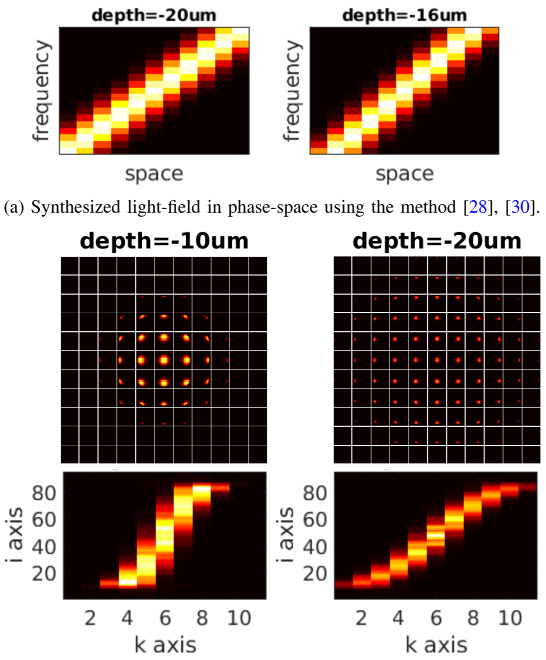
Figure 19. 3D localization results after randomly changing the horizontal and vertical positions. The blue round points represent the reference and the red triangle points represent the detection results. (a)-(d): for non-scattering case. The average RMSE of 3D localization is 2.06 µm, 2.12 µm and 2.39 µm for x, y and z positions, respectively. (e)-(h): for scattering case. The average RMSE of 3D localization is 1.73 µm, 2.00 µm, 1.71 µm for x, y and z positions, respectively.



(a) Dictionary for ideal point sources (b) Dictionary for ball-shaped sources

(c) RMSE of sparse representation

Figure 20. Impact of dictionary modelling. (1) CSC results using a PSF dictionary which models ideal point sources. Blue represents CSC coefficients. Red represents selected coefficients via clustering. (2) CSC results using our more realistic dictionary which models ball-shaped sources. Our dictionary model gives sparser, cleaner and more structured coefficients. (3) CSC with respect to our dictionary model leads to more accurate sparse representation with smaller RMSE (red curve) than using the PSF dictionary (blue curve).



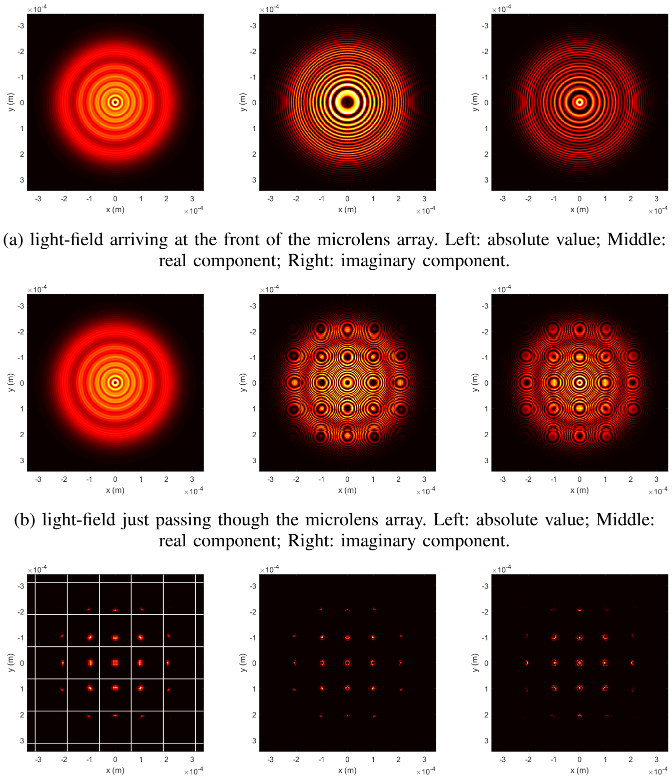
(a) Synthesized light-field in phase-space using the method [28], [30].

(b) Synthesized light-field in image space and EPI space using our approach.

Figure 21. Comparing light-fields synthesized using the method [28], [30] and our method.

Note that the footprint dictionary in [28] is constructed from real data. If the tissue sample is changed, a new dictionary needs to be constructed. In contrast, the dictionary in our paper is composed of a set of EPIs, each corresponding to a specific depth. The convolutional sparse coding with respect to the dictionary gives coefficients which correspond to the 3D positions. Specifically, the identified EPI atoms directly give the depths. The construction of our dictionary does not require real data. It does not need to be changed thereafter when applied to other tissue samples of the same type. Moreover, comparing with [28], [30], we do not need to incorporate light scattering in the dictionary, since we introduce a purification operation based on matrix factorization to mitigate the blurring due to scattering.

2) Optical Aberration: Optical aberration has been considered in the design of the microscopy system, in the mathematical modelling, and in experiment implementation. During the design of LFM system, we have tried to correct optical aberrations in order to create an optimal system. In particular, we used achromatic lenses which are designed for a range of wavelengths to limit the effects of chromatic and spherical aberration. In addition, we also tried to minimise monochromatic aberrations by aligning the microscope with the wavelength emitted from the fluorophore. Such calibrations



(a) light-field arriving at the front of the microlens array. Left: absolute value; Middle: real component; Right: imaginary component.

(b) light-field just passing though the microlens array. Left: absolute value; Middle: real component; Right: imaginary component.

(c) light-field arriving at the sensor plane. Left: absolute value; Middle: real component; Right: imaginary component.

Figure 22. Simulating light-field along the whole path of propagation for a point source located at depth of 10 μm. (a) shows the light-field that has passed through the main lens and 4f-system, arrives at the front of the microlens array. (b) shows the light-field just passing through the microlens array. (c) shows the light-field arriving at the sensor plane. The white grid represents the profile of the microlens array.

made acquired light-field image clear and sharp.

Furthermore, during modelling, instead of exploiting ray optics, we leveraged wave-optics [24] to simulate light-field along the whole path of propagation. This also takes aberrations caused by the wave nature of light into consideration and helps to ensure the faithfulness of our model to the real observations. We note in Fig. 22 the airy disk diffraction patterns caused by circular aperture of main lens and they contain typical aberrations of a diffraction-limited systems such as defocus aberrations.

Finally, in the localization experiment, we adopt matrix-factorization based post-processing operation, convolutional sparse coding and clustering techniques to enhance the robustness of our approach to aberrations, scattering and other latent interference.

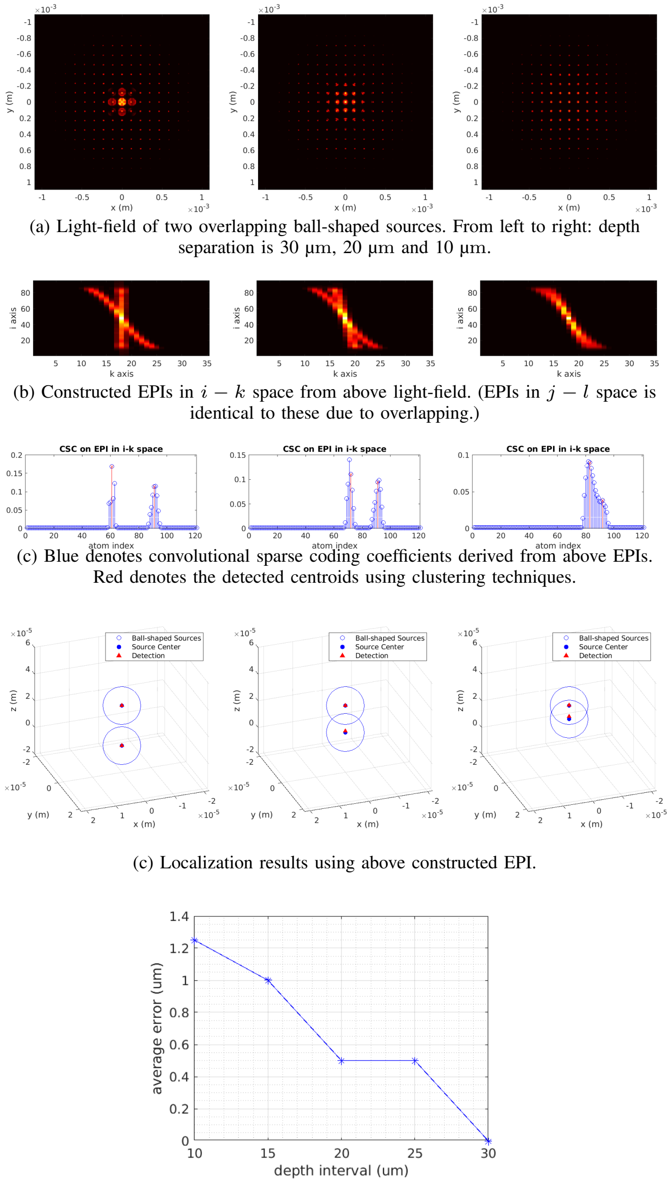
However, we also need to point out that an optical system may involve various aberrations, such as coma, astigmatism, field curvature, etc. To fully incorporate their effects, it requires to model these aberrations explicitly, for example, by adding an appropriate Zernike polynomial to the wavefront in k-space. Such sophisticated modelling would make the model over complicated and this may impact the efficiency of our approach. Given that our current model has already provided satisfactory performance, we prefer to leave such research to our future work.

3) Scattering: Even though the amount of scattering can be estimated and incorporated into the forward model during the dictionary construction, it still hamper the localization performance. This fact has also been noted in other works [28], [30]. It is noticed that the scattering from the tissue sample is nearly homogeneous across the entire image region and exhibits similar pattern at various neutral structures. Therefore, it acts as an impediment to localization of neurons, rather than facilitating the task. That accounts for why we resort to alternative ways to mitigate the impact of scattering.

In particular, it is observed that tissue scattering follows a specific distribution, similar to perlin noise, which provides different patterns from the neurons. By capitalizing on the redundant information contained in a set of multi-view images, such specific patterns can be effectively separated using matrix factorization techniques. The validity of this method has also been verified in our experiments. It shows that the scattering can be effectively distinguished and then subtracted from the target objectives in the foreground. Accordingly, neurons are revealed more clearly in the purified multi-view images. Even though it may not be possible to eliminate the scattering completely, the impact of the remaining scattering is marginal due to algorithm's robustness induced by sparsity.

Even though the proposed purification method empirically works well, it also has a risk of failure, for example, excluding a deep source if it is completely covered by the scattering. Since all the structures have been submerged, performing localization in such a circumstance is highly challenging for other methods as well. The proposed purification operation can be successfully applied to a depth range of $[-40, 40]$μm according to our experiments with brain samples of a genetically encoded mouse.

4) Sample Density: For the non-scattering case, the sample density at the x-y transverse plane is subject to transverse resolution. In particular, the resolution limit at the sample plane is derived using the lenslet pitch and magnification factor, i.e. $d/M = 125\mu\text{m}/25 = 5\mu\text{m}$, where d is the lenslet pitch and $M = 25\times$ is the magnification factor. Therefore, the transverse resolution is 5 μm, which suggests that two adjacent ideal point sources whose distance is less than 5 μm can not be distinguished, as their epipolar lines admit the same intercept with angular axis i and j. Furthermore, considering the diameter, i.e. 10 μm, of ball-shaped volumetric sources, the resolvable separation distance in x-y plane reduces to 15 μm. The sample density in the depth axis is subject to the radius of ball-shaped sources and the granularity of the EPI dictionary. If the granularity of the EPI dictionary is poor, for example, two adjacent EPI atoms representing 50 μm depth separation, two sources with the depth separation smaller than that value cannot be successfully distinguished. Given an EPI dictionary with sufficiently fine granularity, our experiment demonstrates that the minimum depth separation that still allows successful localization of two adjacent overlapping ball-shaped sources is equal to the source diameter, that is, 10 μm in our case. A depth separation smaller than this results in degraded epipolar lines with large overlap, and this makes it challenging to separate two sources, as shown in Fig. 23. For scattering cases, the resolution may become worse due to



(a) Light-field of two overlapping ball-shaped sources. From left to right: depth separation is 30 µm, 20 µm and 10 µm.

(b) Constructed EPIs in $i-k$ space from above light-field. (EPIs in $j-l$ space is identical to these due to overlapping.)

(c) Blue denotes convolutional sparse coding coefficients derived from above EPIs. Red denotes the detected centroids using clustering techniques.

(c) Localization results using above constructed EPI.

(c) Localization error with respect to different depth separations.

Figure 23. Impact of sample density, in particular, depth separation between two overlapping ball-shaped sources. Two overlapping sources with depth separation less than 10 µm may not be successfully distinguished as their epipolar lines are almost identical. Best seen by zooming on a computer screen.

the blurring introduced by light scattering. However, owing to the proposed purification operation, the blurring can be mitigated considerably. We empirically found that the impact of scattering is marginal. Therefore, we conjecture that the sample density does not deviate too much from the non-scattering case.

5) *Localization Range:* Theoretically, the resolvable depth range depends on the depth of field which is the ability to distinguish features at different depths by refocusing the microscope. It is closely related to angular resolution. We adopt the method proposed in [23] to derive the angular resolution and depth of field. As mentioned in the introduction, microlens-based light-field microscopy has a trade-off between spatial and angular resolution. In a light-field microscopy, the total resolution $N_i \times N_j \times N_k \times N_l$ is limited by the number of resolvable sample spots in the specimen. A commonly used measure of this resolution is the Sparrow limit [49], which is defined as $R_{obj} = \frac{0.47\lambda}{NA}M$ where λ is the wavelength of the light, NA is the Numerical Aperture, and M is the magnification factor. Our LFM system leads to a sparrow limit $R_{obj} = \frac{0.47\lambda}{NA}M = 0.47*0.65*25/1.0 = 7.64\mu\text{m}$ at the imaging side. Given the lenslet pitch $d = 125\mu\text{m}$, the angular resolution is $N_i = N_j = d\times/R_{obj} = 125/7.64 = 16.36$. The depth of field is $DOF = \frac{(2+N_i^2)\lambda n}{2NA^2} = \frac{(2+16.36^2)*0.65*1}{2*1.0^2} = 87.64\mu\text{m}$, where n denotes the refractive index of the material. This implies that if two point sources are separated by a distance grater than DOF along the depth axis, they cannot be simultaneously localized. In practice, the resolvable depth range also depends on the intensity contrast between the image of cells in the foreground and the blurring in the background resulting from the tissue scattering. Such contrast leads to a proportional intensity contrast between epipolar lines and background in an EPI. Deeper sources tend to produce weaker intensity contrast, and this makes it more challenging to distinguish the epipolar lines from blurred background. According to our simulation, for the case of a single source without scattering, the proposed localization algorithm performs well at depths up to 120 µm, as shown in Fig. 24. Since our approach is based on convolution, the estimation of the x-y coordinates is not affected by depth and transverse dimensions.

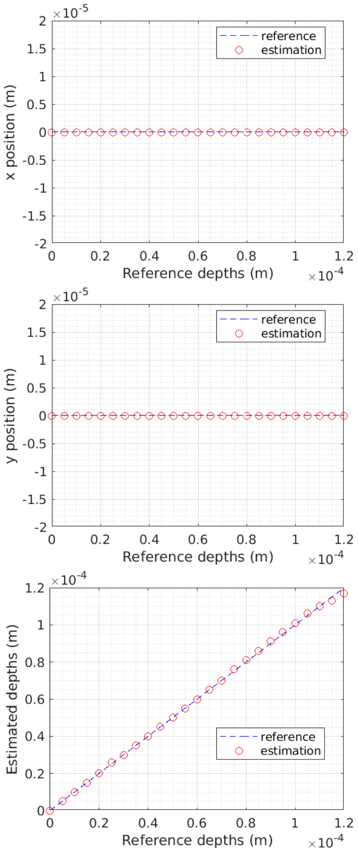
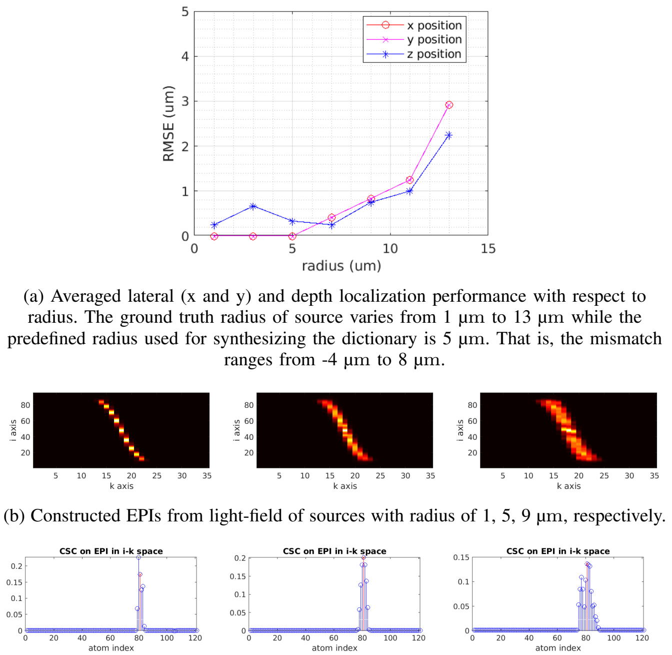


Figure 24. Localization performance with respect to depth range. Depth range is from 0 µm to 120 µm.

6) *Impact of Radius Mismatch:* To investigate the impact of radius of ball-shaped sources on localization performance, we vary the radius of ball-shaped volumetric sources while fixing



(a) Averaged lateral (x and y) and depth localization performance with respect to radius. The ground truth radius of source varies from 1 μm to 13 μm while the predefined radius used for synthesizing the dictionary is 5 μm. That is, the mismatch ranges from -4 μm to 8 μm.

(b) Constructed EPIs from light-field of sources with radius of 1, 5, 9 μm, respectively.

(c) Blue denotes convolutional sparse coding coefficients derived from above EPIs. Red denotes the detected centroids using clustering techniques.

Figure 25. Impact of radius mismatch on localizing a single ball-shaped source. In general, if the ground truth radius r is smaller than the predefined radius r^* used for synthesizing the dictionary, i.e $r < r^*$, the impact is minor. When the ground truth is larger than the predefined one, i.e. $r > r^*$, along with the increase of the mismatch, the localization performance tends to degrade and introduce larger deviation. On the other hand, our approach demonstrates satisfactory robustness at a reasonable mismatch range $r - r^* \in (-5, 5)$ μm. Best seen by zooming on a computer screen.

the predefined radius used for synthesizing the dictionary.

Fig. 25 shows the impact of radius mismatch on localizing a single ball-shaped source. The ground truth radius r varies from 1 μm to 13 μm with a step of 2 μm and the predefined radius r^* used to synthesize the EPI dictionary is kept at 5 μm. Therefore, the mismatch between the estimated radius and the ground truth ranges from -4 μm to 8 μm. In general, if the ground truth radius r is smaller than the predefined radius r^*, the impact is minor. When the ground truth is larger than the predefined one, i.e. $r > r^*$, along with the increase of the mismatch $r - r^*$, the localization performance degrades. However, our approach demonstrates satisfactory robustness at a reasonable mismatch range $r - r^* \in (-5, 5)$ μm. This is due to the fact that within a certain mismatch range, the coefficients found by convolutional sparse coding are still sparse and structured, as shown in Fig. 25 (c). Therefore, the sparsity prior and clustering technique exploited in our algorithm enable us to find a good approximation from the dictionary. Once the mismatch is over a limit, the coefficients are not sparse enough any more and this leads to more significant deviation.

To summarize, the impact of radius mismatch is not significant within a certain range due to the algorithm robustness induced by the sparsity prior and clustering technique exploited in the proposed approach.