Supplementary Information for SparseFocus: Learning-based One-shot Autofocus for Microscopy with Sparse Content

Yongping Zhai[1*†], Xiaoxi Fu[1†], Qiang Su[1†], Jia Hu[1], Yake Zhang[1], Yunfeng Zhou[1], Chaofan Zhang[1], Xiao Li[1], Wenxin Wang[1], Dongdong Wu[2], Shen Yan[3*]

[1]College of Advanced Interdisciplinary Studies, National University of Defense Technology, Changsha, 410073, Hunan, China.
[2]Department of Information, Army Medical University, Chongqing, 400042, Chongqing, China.
[3]College of System Engineering, National University of Defense Technology, Changsha, 410073, Hunan, China.

*Corresponding author(s). E-mail(s): zhaiyongping08@nudt.edu.cn; yanshen12@nudt.edu.cn;
†These authors contributed equally to this work.

S1 Analysis of Sharpness Curve

The sharpness curve characterizes the functional relationship between imaging sharpness and z-position. Ideally, it exhibits three critical properties, as illustrated in Fig. S1:

- Unbiased nature: the maximum sharpness precisely coincides with the focal plane position;
- Unimodality: a single maximum exists, with strict monotonicity in the intervals partitioned by the maximum index;
- Sensitivity: capability to distinguish in-focus images from slightly defocused images.

However, in practical z-stack images, this curve markedly deviates from the ideal morphology due to factors including sample sparsity, thickness, and multi-layer structures.

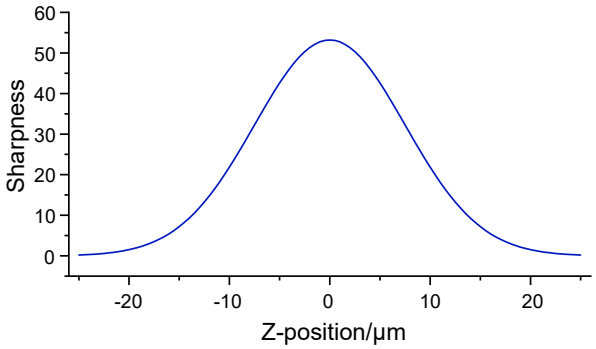

Fig. S1 Sharpness curve under ideal conditions.

S1.1 Sparsity

Sample sparsity induces high-frequency artifacts originating primarily from content-deficient regions including blank backgrounds and dust-contaminated areas, yielding characteristic non-unimodal sharpness curves. Fig. S2 demonstrates that progressive sparsity escalation across dense, sparse, and extremely sparse specimens within identical z-axis ranges correlates with both diminished sharpness magnitude and emergent secondary peaks.

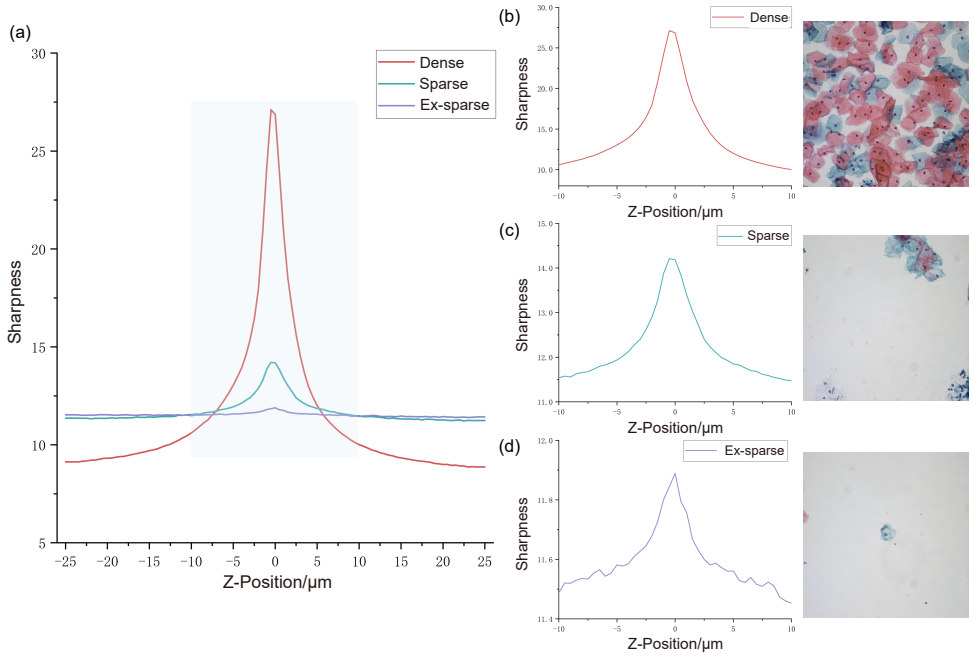

Fig. S2 Sharpness curves and corresponding sample visualizations under dense, sparse, and extremely sparse conditions.

These spurious peaks substantially compromise local optimization-based autofocus algorithms such as hill-climbing methods, inducing convergence toward local maxima rather than the globally optimal focal plane position.

Spatial partitioning analysis in Fig. S3 reveals distinct curve morphologies when segmenting sparse samples into content-rich regions, dust-contaminated zones, and background areas. Crucially, content-rich regions generate unimodal curves resembling those of dense samples, whereas dust and background areas produce noise-dominated traces lacking discernible peaks. This confirms that noise artifacts stem predominantly from dust/background zones, while the attenuated primary peak arises exclusively from content-rich regions.

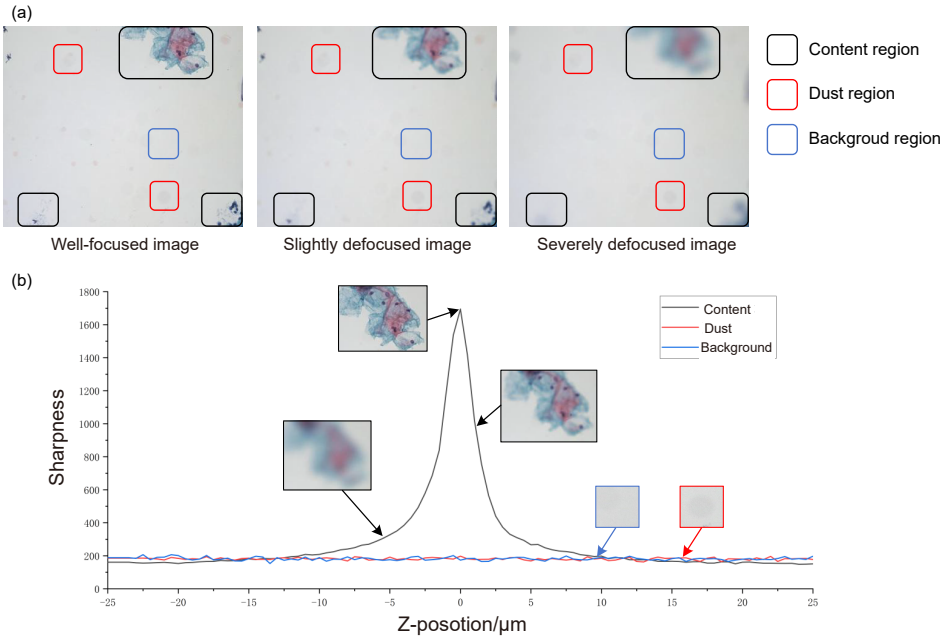

Fig. S3 Sharpness curves across content, dust, and background regions.

S1.2 Thickness

Sample thickness variations induce multiple distinct maxima in sharpness curves, as demonstrated in Fig. S4a. At z-positions corresponding to individual maxima, only partial regions of the image achieve optimal focus while others exhibit significant blurring. Crucially, the focal regions reverse between adjacent maxima – areas previously in focus become blurred and vice versa, illustrated in Fig. S4b.

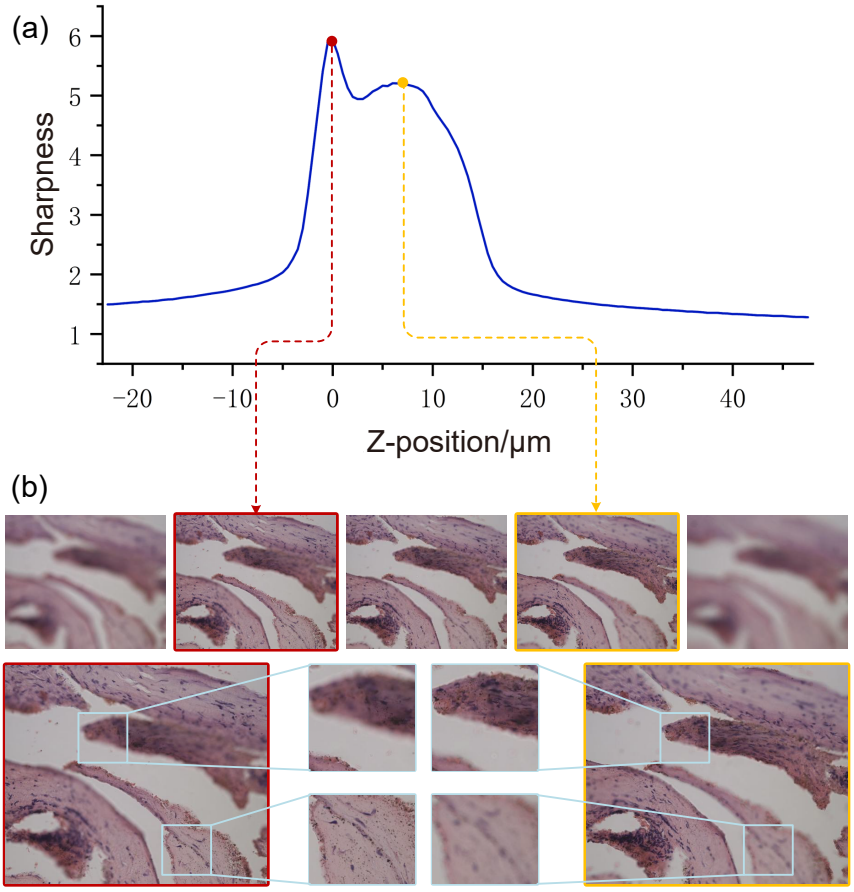

Fig. S4 Sharpness curve and corresponding image details for a thick specimen.

This phenomenon originates from unavoidable thickness heterogeneity during sample preparation, wherein the section plane exhibits non-planar topography. Such variations establish multiple focal planes within a single field of view. When positioned at one focal plane, only specific regions maintain sharp focus while others defocus; conversely, alternative focal planes sharpen previously blurred areas.

As exemplified in a field containing two prominent focal planes, the resulting sharpness curve manifests distinctive bimodal characteristics. Qualitatively, this represents the superposition of two unimodal functions with differing peak positions and magnitudes. To address this challenge, two practical strategies emerge:

- ROI-restricted focusing: Manual selection of subregions for localized focusing
- Z-stack fusion: Computational integration of multiple z-position images to synthesize all-in-focus reconstructions

S1.3 Multi-Layer Structures

Extensive focal search ranges accentuate the influence of multi-layer structures on sharpness curve morphology, as evidenced in Fig. S5a. The first peak corresponds to focus at the coverslip's dust layer interface, while the second peak signifies optimal focus within the specimen layer, as depicted in Fig. S5b.

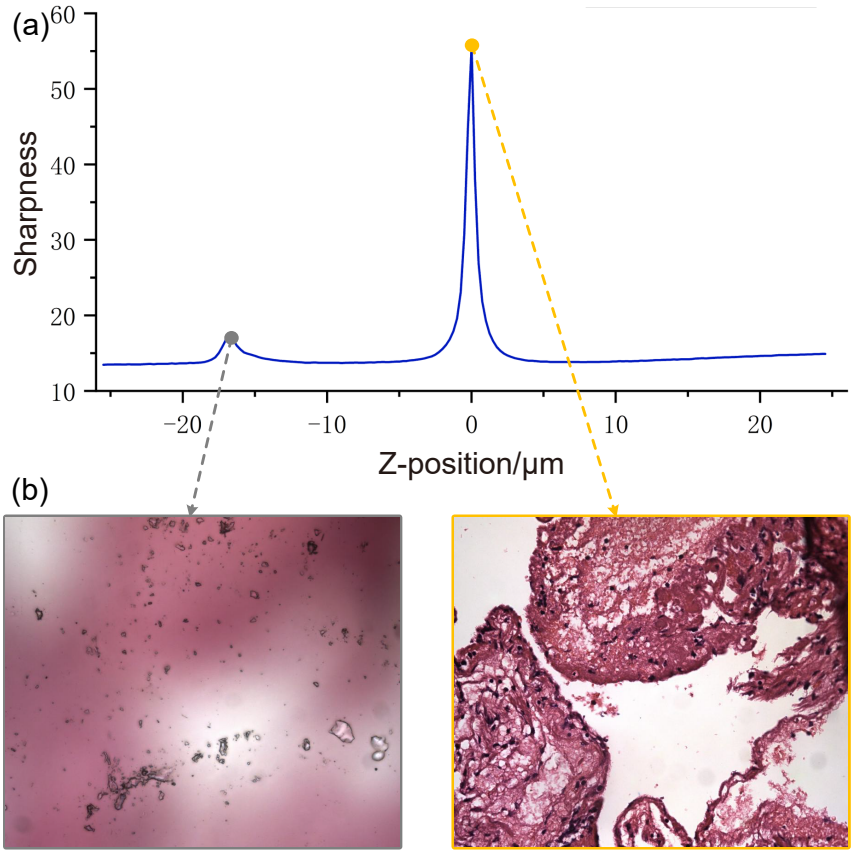

Fig. S5 Sharpness curve and corresponding image details for a sample with multi-layer structures.

Although morphologically analogous to thickness-induced multi-peak curves, hierarchical configurations systematically generate dust-related artifacts. Critically, these artifacts invariably manifest in all coverslip-mounted specimens, necessitating calibrated focusing protocols to eliminate dust layer interference.

S2 Hill-climbing Autofocus Method

Hill climbing autofocus represents a classic passive search algorithm rooted in local optimization strategies and demonstrates favorable accuracy under general conditions. However, its convergence speed is notably slow, typically requiring several seconds to reach the focal plane, due to its dependency on repeated mechanical movements and directional changes during the focus-seeking process. As depicted in Fig. S6, achieving

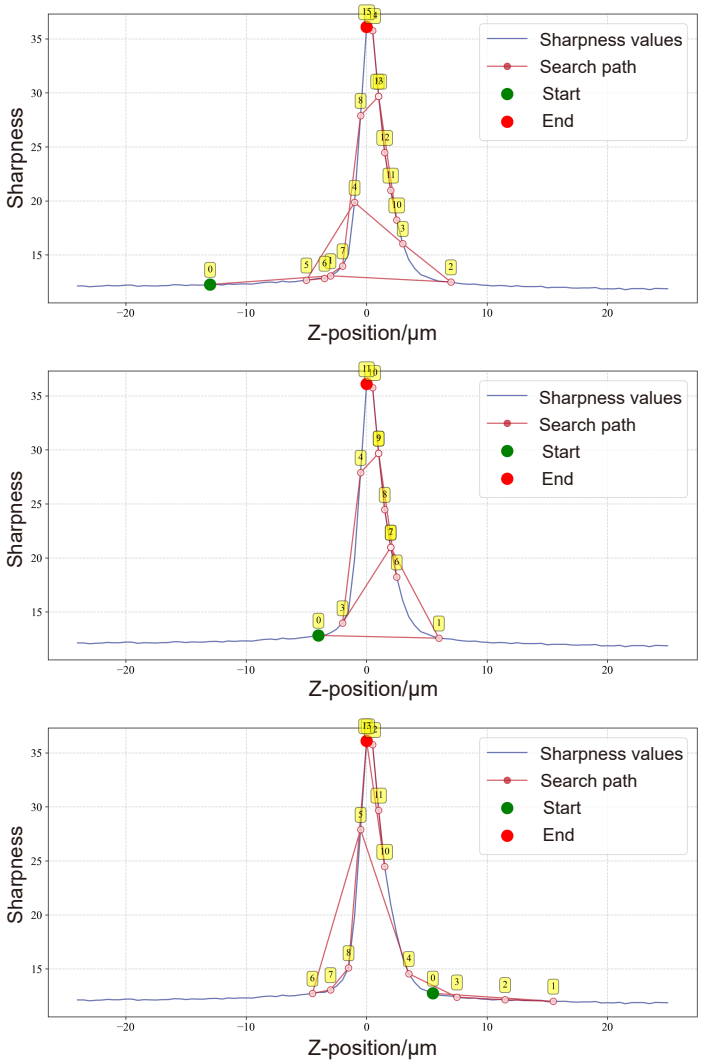

Fig. S6 Sharpness curve characteristics and hill-climbing autofocus process on dense samples.

focus via this method necessitates, on average, over a dozen movements accompanied by multiple directional reversals, rendering it inadequate for real-time requirements in rapid scanning applications.

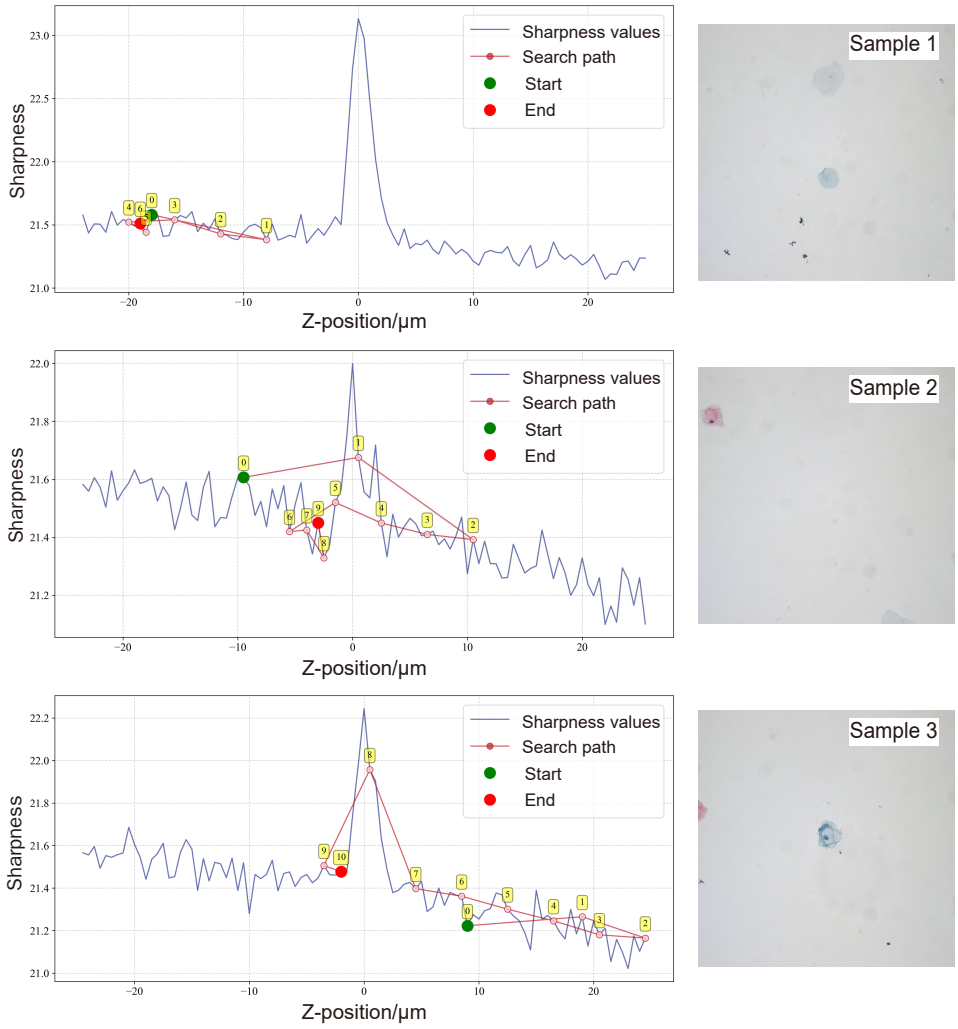

Fig. S7 Sharpness curves, hill-climbing search paths, and in-focus images for three extremely sparse samples.

Beyond its sluggish performance, hill-climbing autofocus exhibits significant susceptibility to sample sparsity. As previously analyzed, within sparse samples, regions characterized by low informational content (e.g., blank background areas, dust artifacts) are prone to high-frequency noise interference. This phenomenon introduces prominent non-unimodality into the focus curve. These secondary peaks critically undermine hill-climbing and other local optimization-based passive autofocus techniques. They divert convergence towards local extrema induced by noise artifacts rather than the global maximum, ultimately causing focus failure. Fig. S7 illustrates this outcome across three distinct sparse to highly sparse scenarios: the convergence point frequently resides on secondary peaks instead of the true focal plane. These secondary peaks exhibit random distribution, with some displaced by tens of micrometers from the in-focus position. When success is defined as convergence to the global maximum, the hill-climbing algorithm attained success rates of only 32.1%, 16.05%, and 30.86% in these respective scenarios.

Table S1 summarizes the comparative results. Across these scenarios, the proposed method significantly outperforms hill-climbing autofocus in both accuracy and computational speed.

Table S1 Comparison of our method and the hill-climbing algorithm on three specific extremely sparse samples in terms of MAE and computational time.

MAE (μm) / Time (ms)	Hill-climbing	Ours
Sample 1	12.02 / 290.01	0.54 / 8.04
Sample 2	11.20 / 248.30	0.63 / 7.57
Sample 3	7.24 / 285.46	0.59 / 9.29

S3 Details of Labeling Content Importance Matrix

Analysis of sharpness curves reveals that in sparse samples, regions such as dust artifacts and background areas predominantly exhibit noise-corrupted characteristics, containing minimal features relevant to defocus distance prediction. Beyond theoretical rationale, extensive experimental evidence demonstrates that even well-trained defocus prediction networks produce near-random predictions for these noise-dominated regions, introducing substantial error into final defocus distance estimates. Consequently, image regions exhibit heterogeneous contributions to defocus prediction: content-rich areas contain meaningful defocus features and higher importance weights, while dust and background regions contribute lower importance with minimal predictive value.

However, identifying regions of high predictive importance traditionally requires sharpness curves or, at minimum, comparative analysis of in-focus and defocused image pairs to determine regional significance. This poses a fundamental limitation for one-shot autofocus applications, where only a single defocused image is available. Conventional image processing techniques cannot accurately identify high-importance

regions under such constraints. To address this, we leverage the powerful feature extraction capabilities of neural networks. Our proposed Region Importance Network (RIN), trained on large datasets, demonstrates the ability to estimate per-region importance directly from a single defocused image - a capability empirically validated in our experiments.

Training such networks requires large volumes of accurately annotated data. Manual annotation based on subjective judgment or sharpness curve guidance is prohibitively inefficient for large-scale applications. We therefore propose an automated annotation framework leveraging an objective criterion derived from sharpness curve analysis: Regional importance correlates positively with the ratio of in-focus sharpness to defocused sharpness. This annotation process is formalized as:

$$CI_C^{(i,j)} = \mathbb{I}\left(\frac{\mathcal{S}\left(\mathbf{I}_{\text{infocus}}^{(i,j)}\right)}{\mathcal{S}\left(\mathbf{I}_{\text{defocus}}^{(i,j)}\right) + \epsilon} > \mathcal{T}\right), \tag{1}$$

where CI_C denotes the sharpness-derived content importance matrix; $CI_C^{(i,j)}$ represents the importance value for region (i, j); $\mathbb{I}(\cdot)$ is the indicator function; $\mathcal{S}(\cdot)$ is the sharpness evaluation function (in this study, we use Laplacian sharpness evaluation function); $\mathbf{I}_{\text{infocus}}^{(i,j)}$ denotes the in-focus image region; $\mathbf{I}_{\text{defocus}}^{(i,j)}$ represents the corresponding defocused region; $\epsilon = 10^{-6}$ prevents division by zero. Empirical observation determined that a threshold $\mathcal{T} = 1.2$ reliably distinguishes content-rich regions from dust/background areas.

For color images, regional saturation provides auxiliary information. A secondary criterion is applied:

$$CI_S^{(i,j)} = \mathbb{I}\left(\frac{\sum_{i \in \text{Sat}\left(\mathbf{I}_{\text{infocus}}^{(i,j)}\right)} \mathbb{I}(i > \mathcal{T}_1)}{\text{Size}\left(\mathbf{I}_{\text{infocus}}^{(i,j)}\right)} > \mathcal{T}_2\right), \tag{2}$$

where CI_S denotes the saturation-derived content importance matrix; $\text{Sat}(\cdot)$ extracts the saturation channel; $\text{Size}(\cdot)$ computes pixel count. Thresholds are empirically set to $\mathcal{T}_1 = 12$ and $\mathcal{T}_2 = 0.1$. The final Content Importance matrix CI combines both criteria via element-wise multiplication:

$$\mathbf{W}^* = CI_C \odot CI_S. \tag{3}$$

Fig. S8 illustrates this comprehensive annotation pipeline. For grayscale images lacking saturation data, the sharpness-derived matrix CI_C serves as the final $\mathbf{W}^*$ matrix.

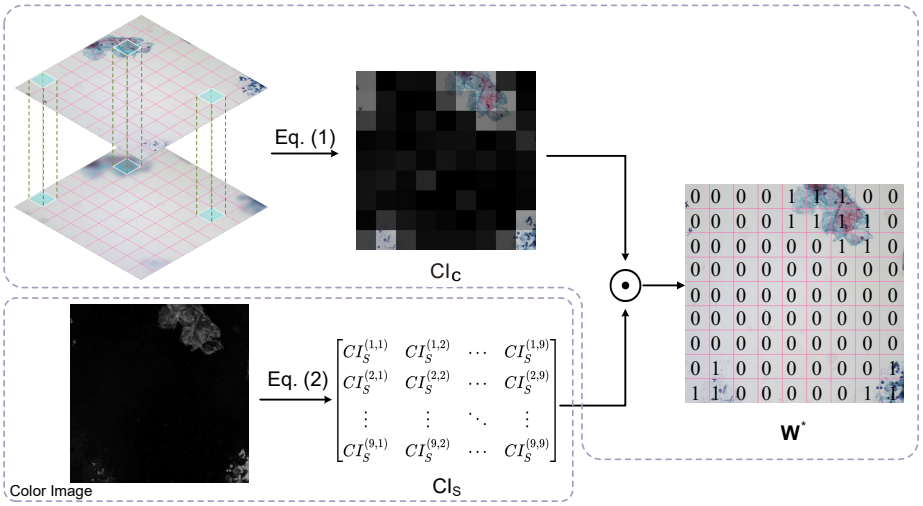

Fig. S8 Schematic of the Content Importance Matrix $\mathbf{W}^*$ Annotation Workflow.

S4 Details of Patch Number Selection

During inference, only the top-k patches are processed by the DPN. This selective strategy confers dual advantages: eliminating low-importance patches enhances prediction robustness while substantially reducing computational load. Extensive experiments conducted on cell samples reveal the selected patch number-MAE relationship as shown in Fig. S9. For extremely sparse samples, the MAE falls below half the DoF when the selected patch number reaches ≥ 3. Thus, $k = 3$ suffices under extreme sparsity. For sparse and dense samples, MAE decreases significantly until $k \leq 9$, plateauing thereafter. At $k = 9$, MAE is less than half the DoF, establishing this value as the optimal trade-off between accuracy and efficiency.

Notably, larger k values do not guarantee improved accuracy. Comprehensive results conducted on both cell and tissue samples in Fig. S10 demonstrate tissue samples exhibit MAE degradation beyond $k \geq 31$. When prioritizing precision over speed in dense scenarios, we recommend $k \leq 31$.

The final number of selected patches is determined by the following:

$$N = \min\left(\sum_{w_{i,j} \in \widehat{\mathbf{W}}} \mathbb{I}(w_{i,j} \geq \rho), k\right), \tag{4}$$

where $\widehat{W}$ denotes the content importance prediction matrix, $\rho = 0.8$ is the importance threshold, and k values follow the experimentally determined constraints above.

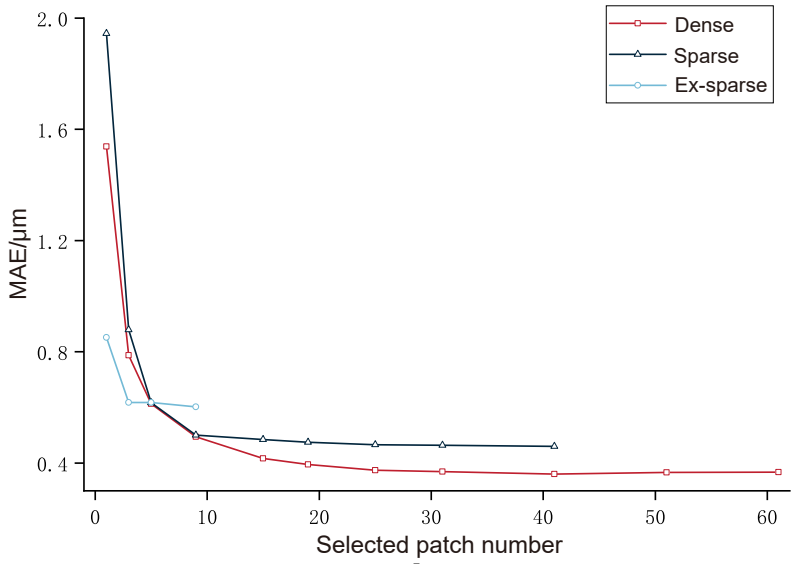

Fig. S9 MAE vs. selected patch number across dense, sparse, and extremely sparse cell samples.

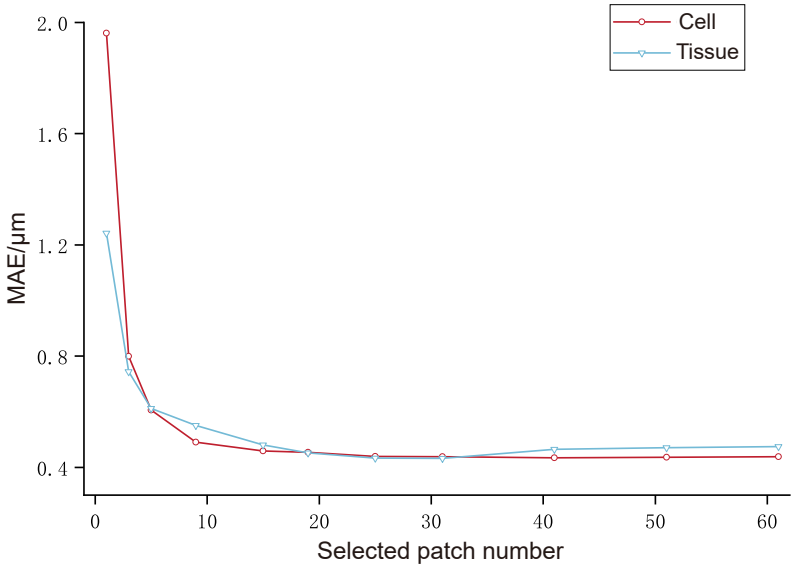

Fig. S10 MAE vs. selected patch number on cell and tissue samples respectively.

S5 More about PSF in Microscopy

The Point Spread Function (PSF) is a theoretical model that describes the response of an optical imaging system to a point source. It represents the light field distribution

of the image of an ideal point-like object and is a primary parameter for evaluating various microscopic imaging systems.

In microscopic imaging of biological samples, optical aberrations, especially spherical aberrations, arise due to mismatches in the refractive indices of the sample, cover slip, and immersion medium. This leads to asymmetry in the PSF on either side of the focal plane. According to the analysis by Gibson and Lanni, the PSF in three-dimensional space can be represented by the Equation (5).

$$
\begin{aligned}
h(x,y,z) &= \left| \frac{C}{z_d} \int_0^1 J_0 \left[ka\rho \frac{\sqrt{x^2+y^2}}{z} \right] e^{jW(\rho)} \rho d\rho \right| \\
W(\rho) &= k \left[n_s t_s \sqrt{1 - \left(\frac{NA \cdot \rho}{n_s} \right)^2} + n_g t_g \sqrt{1 - \left(\frac{NA \cdot \rho}{n_g} \right)^2} + \right. \\
& \left. n_i t_i \sqrt{1 - \left(\frac{NA \cdot \rho}{n_i} \right)^2} - n_g^* t_g^* \sqrt{1 - \left(\frac{NA \cdot \rho}{n_g^*} \right)^2} - n_i^* t_i^* \sqrt{1 - \left(\frac{NA \cdot \rho}{n_i^*} \right)^2} \right],
\end{aligned}
\tag{5}
$$

where (x, y, z) represents the coordinate of a point in the image plane, $h(x, y, z)$ corresponds to the intensity of the light field at this point, C represents a normalizing constant, z_d corresponds to the axial distance between the detector and the tube lens, which is assumed to be equal to the focal length of the tube lens, J_0 represents the Bessel function of first kind of order zero, k corresponds to wavenumber in vacuum of the emitted light, ρ represents radial distance from the light source to the imaging plane, $W(\rho)$ corresponds to the Gibson and Lanni phase aberration, NA represents Numerical Aperture of the objective lens. n_s, n_g and n_i respectively corresponds to refractive index of the sample layer, the coverslip and the immersion layer, t_s represents axial location of the point source within the sample layer, t_g corresponds to actual value of the coverslip thickness, t_i represents actual working distance between the objective and the coverslip, n_g^* and t_i^* represents the corresponding nominal value.

S6 Details of Dataset

Fig. S11(a) showcases a selection of Whole Slide Imaging (WSI) samples from our dataset. The samples exhibit varying degrees of content sparsity, including examples of extremely sparse content highlighted within orange boxes. Fig. S11(b) presents z-stack images that demonstrate symmetry around the focal plane, featuring images at different defocus distances. Fig. S11(c) illustrates the range of content density—dense, sparse, and extremely sparse—visible in each field of view within the microscope for cells and tissues, respectively. Fig. S12 displays representative sample images with defocus distances ranging from -25 µm to 25 µm, encompassing both cell and tissue samples across dense, sparse, and extremely sparse conditions.

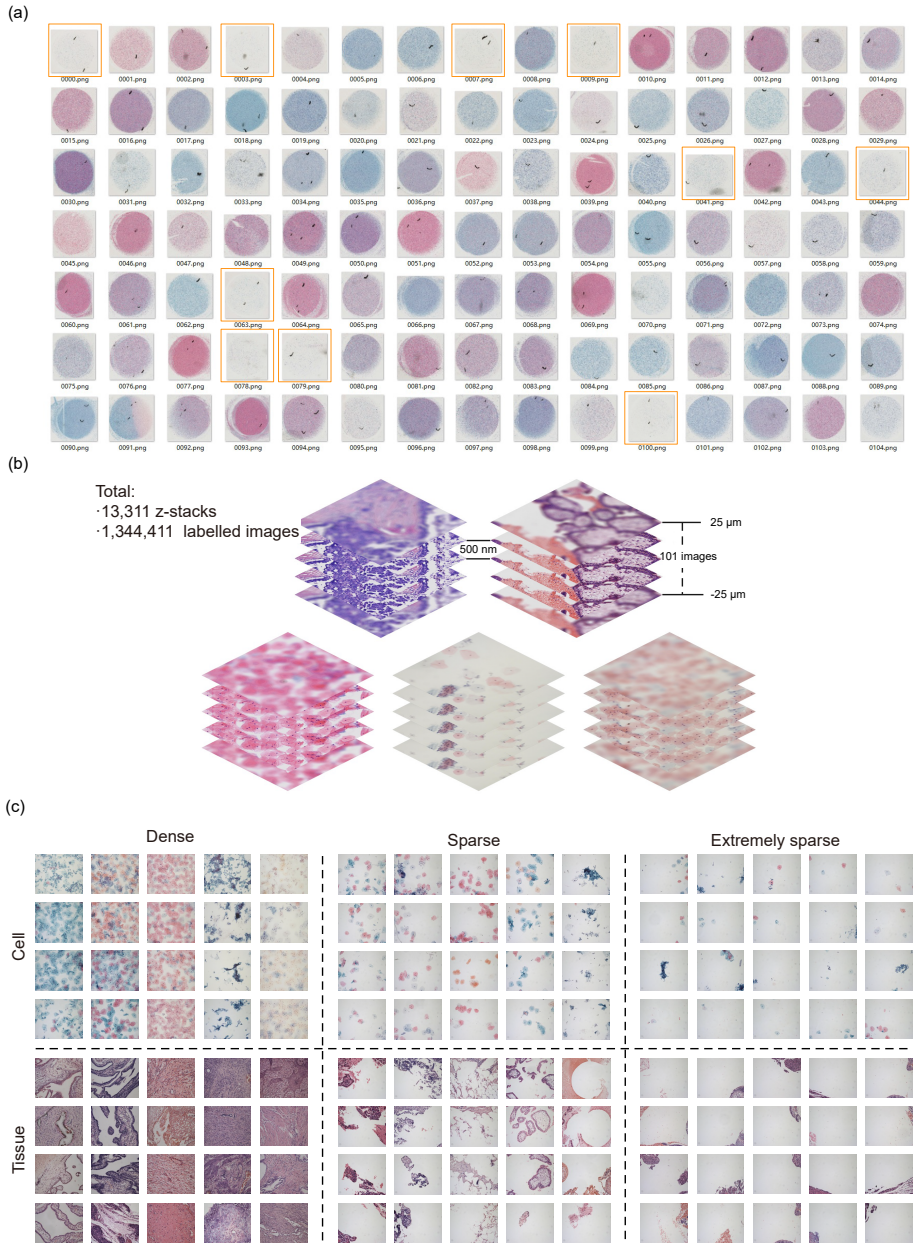

Fig. S11 Representative samples from our large-scale dataset. (a) The displays of Whole Slide Imaging (WSI) samples, with areas marked by orange boxes indicating extremely sparse cases. (b) The top row displays images of pathological tissue, while the bottom row presents images of cervical exfoliated cells. Each sample is captured as a sequence of z-stack images, spanning a defocus range from −25μm to +25μm, with intervals of 500nm. (c) Both cell and tissue categories are represented across three sparsity levels: dense, sparse, and extremely sparse.

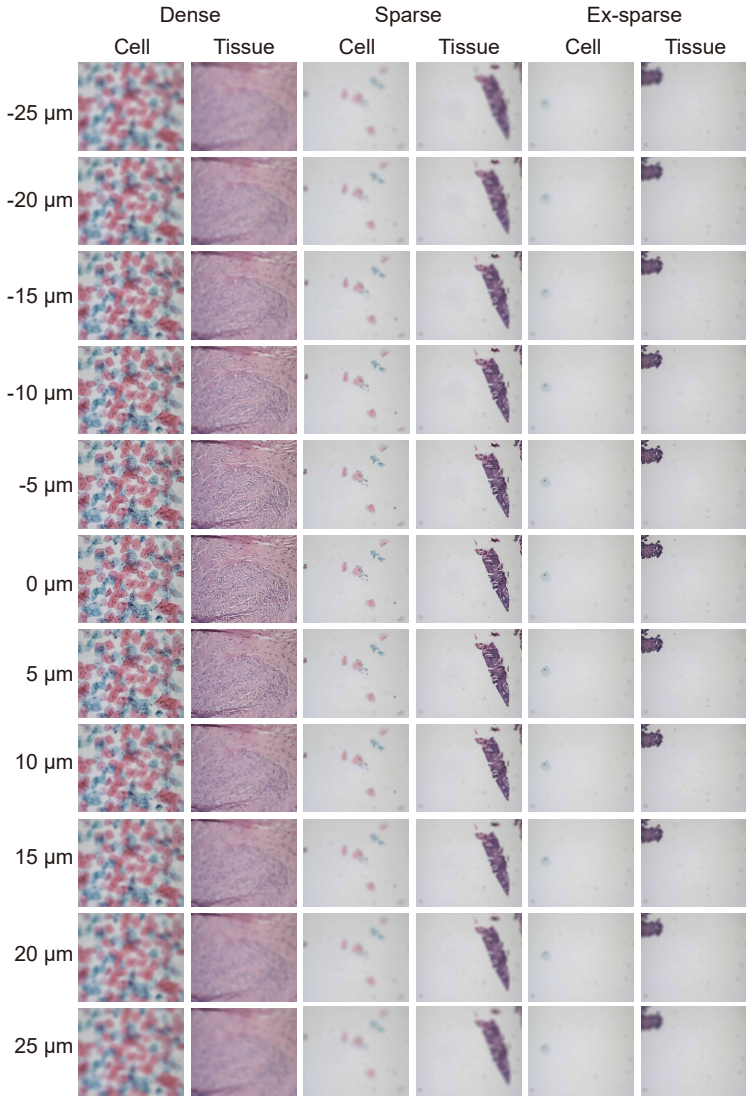

Fig. S12 Representative detailed sample images from our large-scale dataset. These images illustrate defocus distances ranging from −25µm to +25µm, encompassing both cell and tissue specimens across dense, sparse, and extremely sparse conditions.

S7 More Experimental Results

In this section, we present more regression plots of our experimental results.

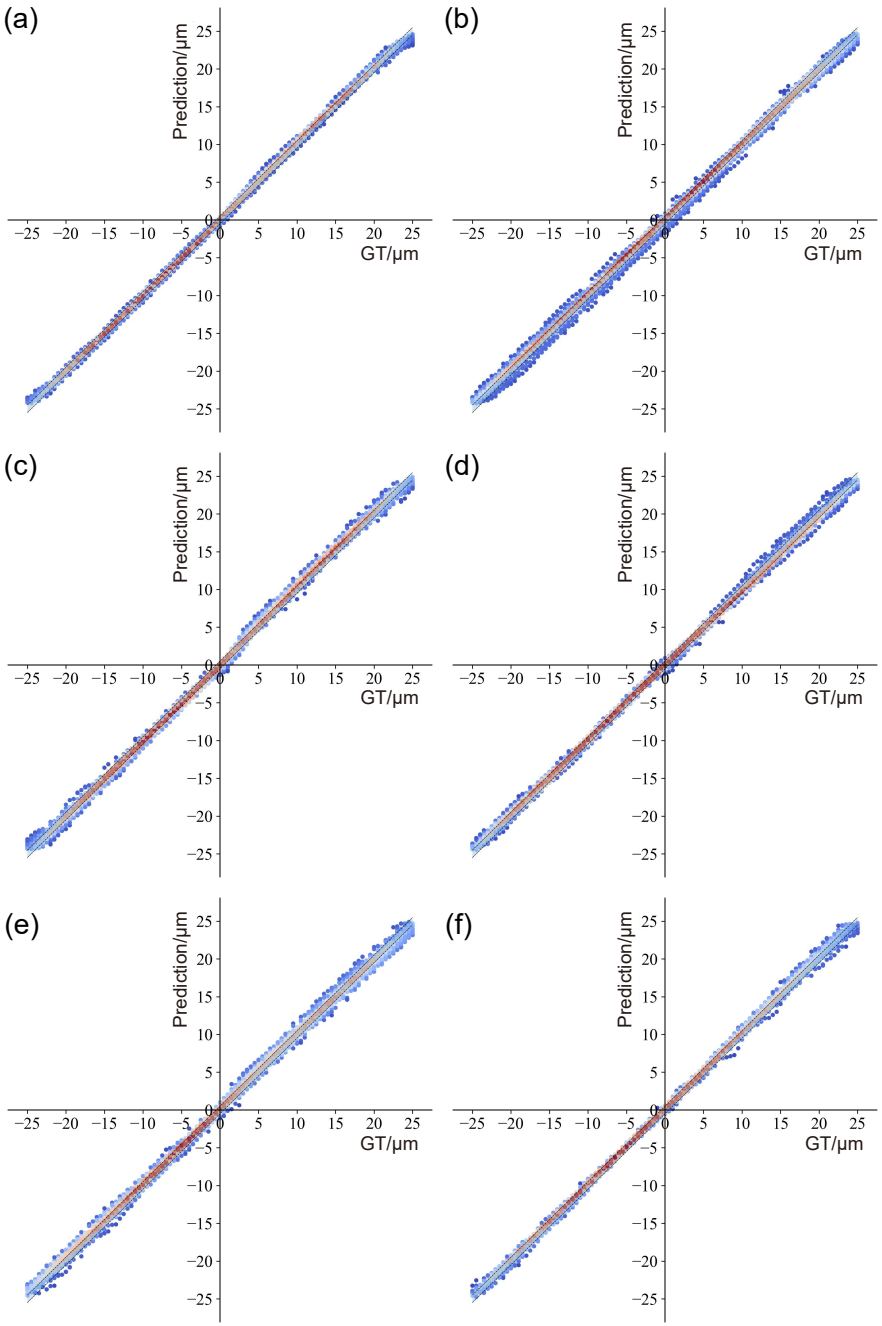

Fig. S13 Regression plots of ours (SparseFocus). (a) shows the results on cell samples in dense scenarios, (b) shows the results on tissue samples in dense scenarios, (c) shows the results on cell samples in sparse scenarios, (d) shows the results on tissue samples in sparse scenarios, (e) shows the results on cell samples in extremely sparse scenarios, (f) shows the results on tissue samples in extremely sparse scenarios.

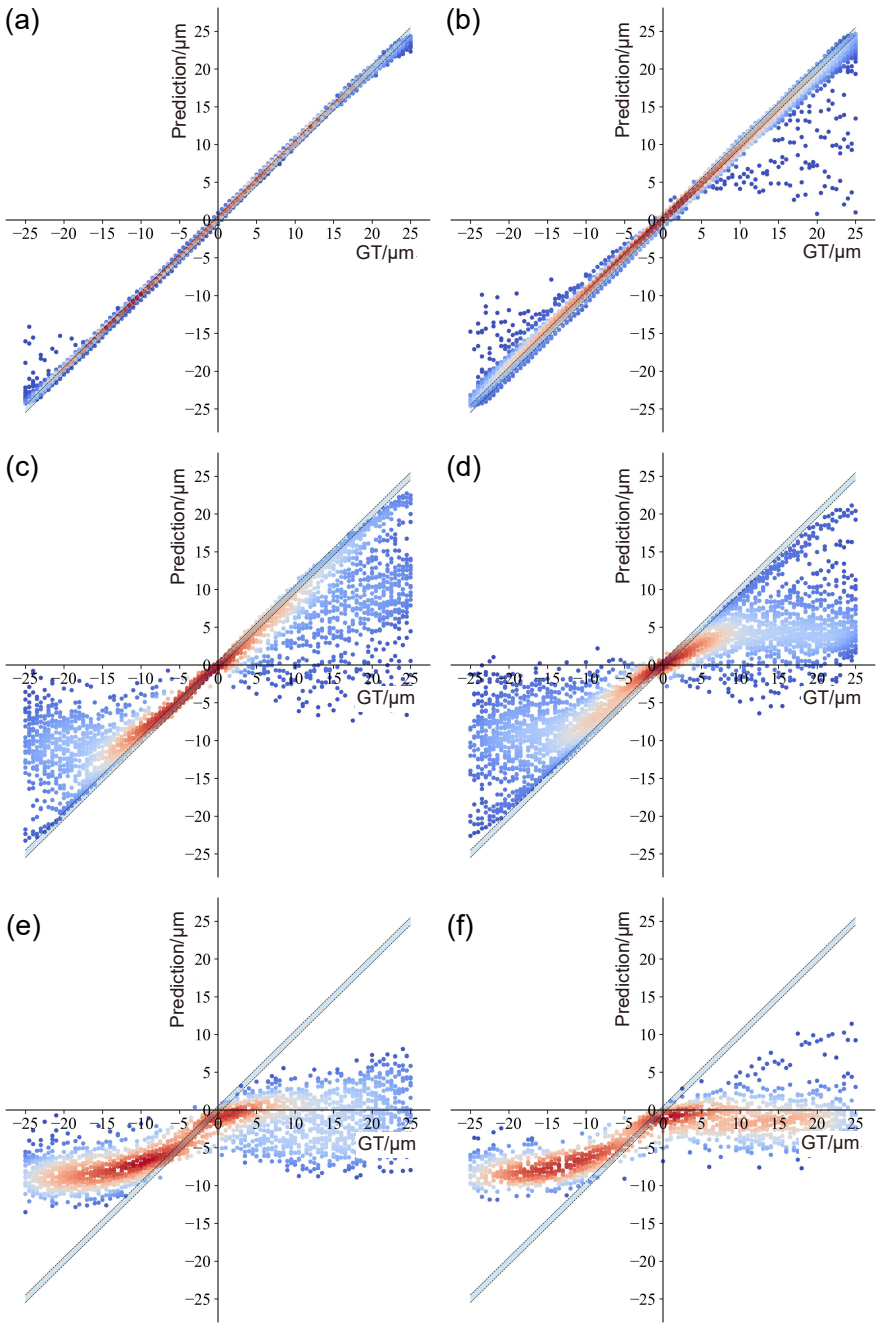

Fig. S14 Regression plots of Dastidar et al. [1]. (a) shows the results on cell samples in dense scenarios, (b) shows the results on tissue samples in dense scenarios, (c) shows the results on cell samples in sparse scenarios, (d) shows the results on tissue samples in sparse scenarios, (e) shows the results on cell samples in extremely sparse scenarios, (f) shows the results on tissue samples in extremely sparse scenarios.

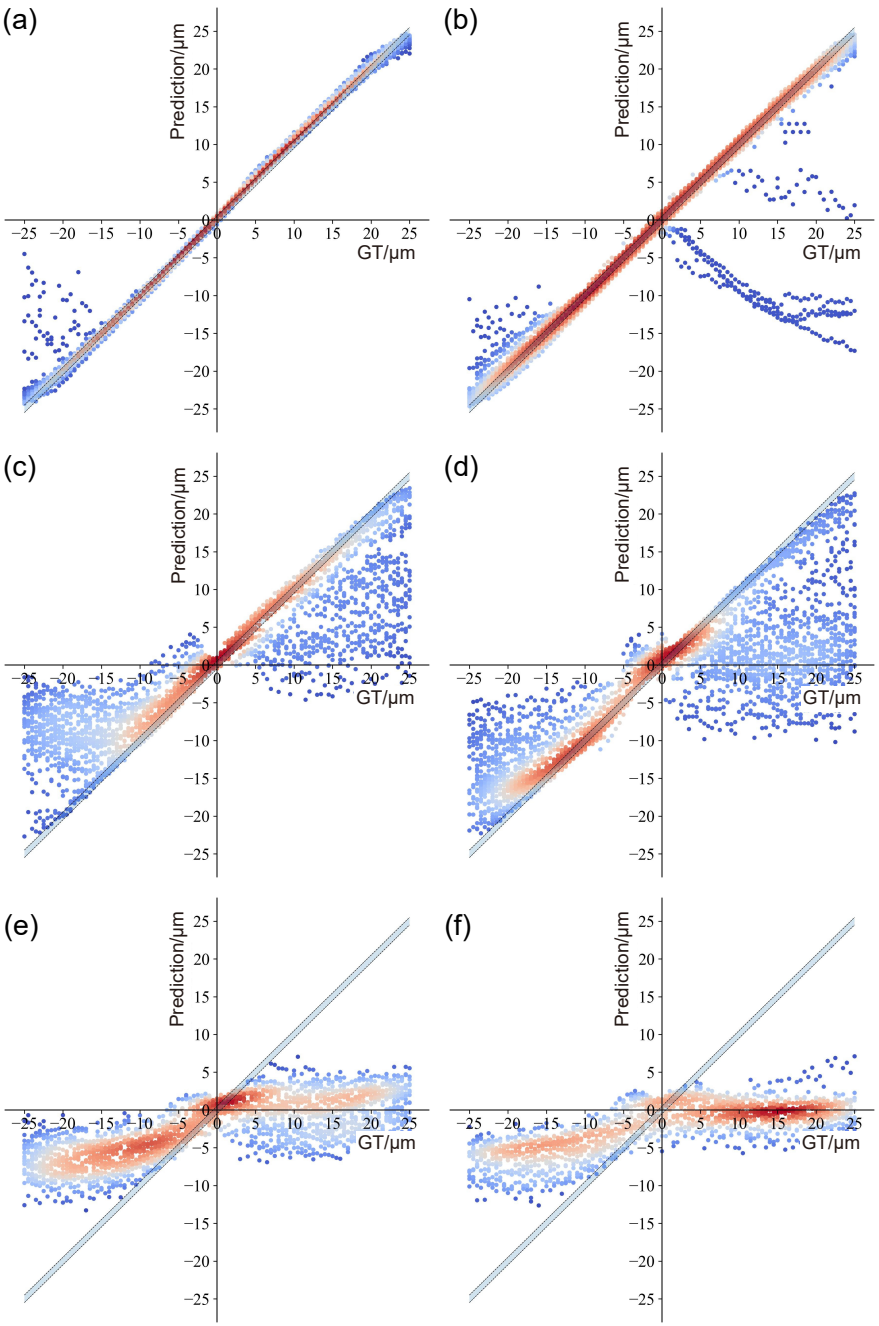

Fig. S15 Regression plots of Liao et al. [2]. (a) shows the results on cell samples in dense scenarios, (b) shows the results on tissue samples in dense scenarios, (c) shows the results on cell samples in sparse scenarios, (d) shows the results on tissue samples in sparse scenarios, (e) shows the results on cell samples in extremely sparse scenarios, (f) shows the results on tissue samples in extremely sparse scenarios.

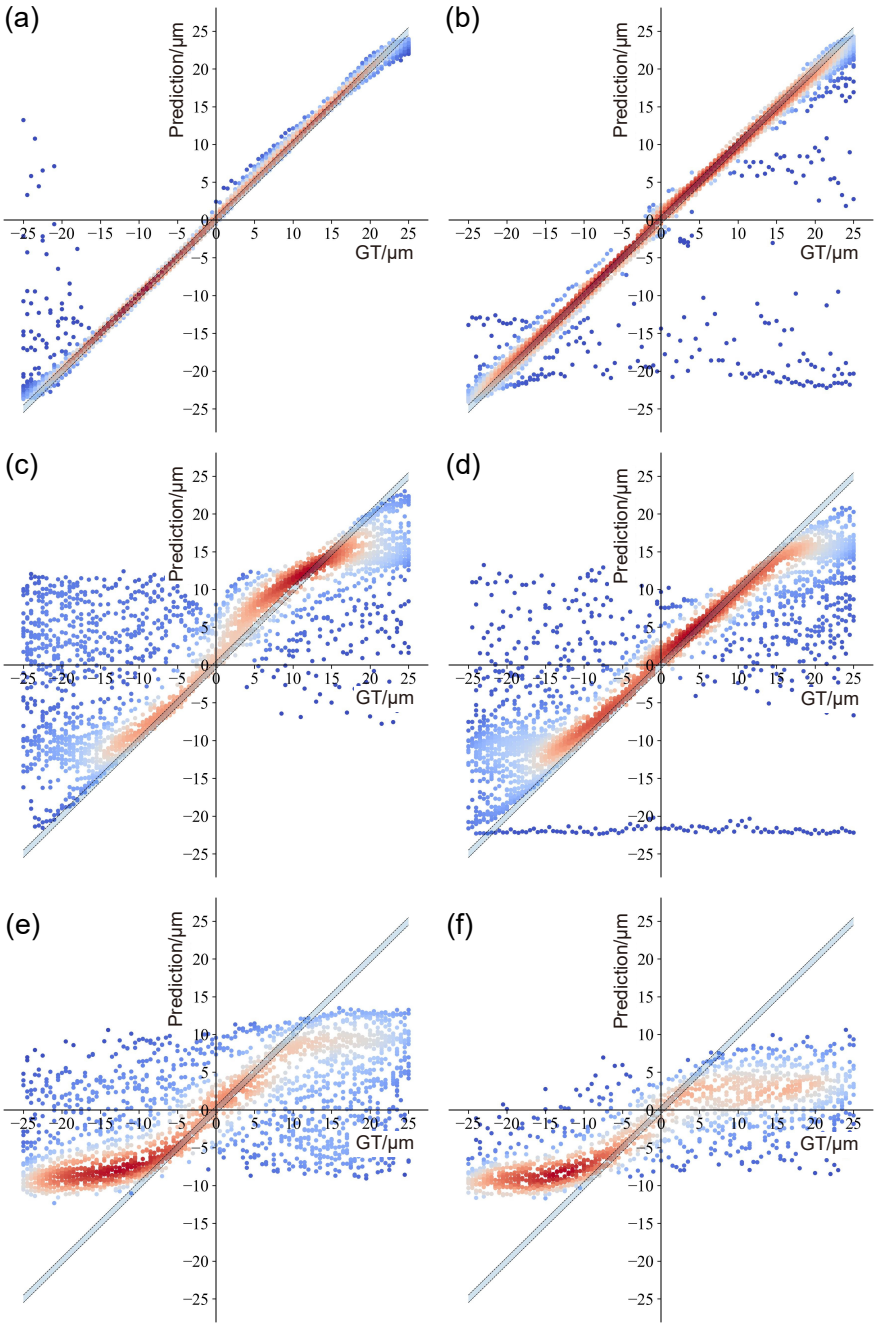

Fig. S16 Regression plots of Jiang et al. [3]. (a) shows the results on cell samples in dense scenarios, (b) shows the results on tissue samples in dense scenarios, (c) shows the results on cell samples in sparse scenarios, (d) shows the results on tissue samples in sparse scenarios, (e) shows the results on cell samples in extremely sparse scenarios, (f) shows the results on tissue samples in extremely sparse scenarios.

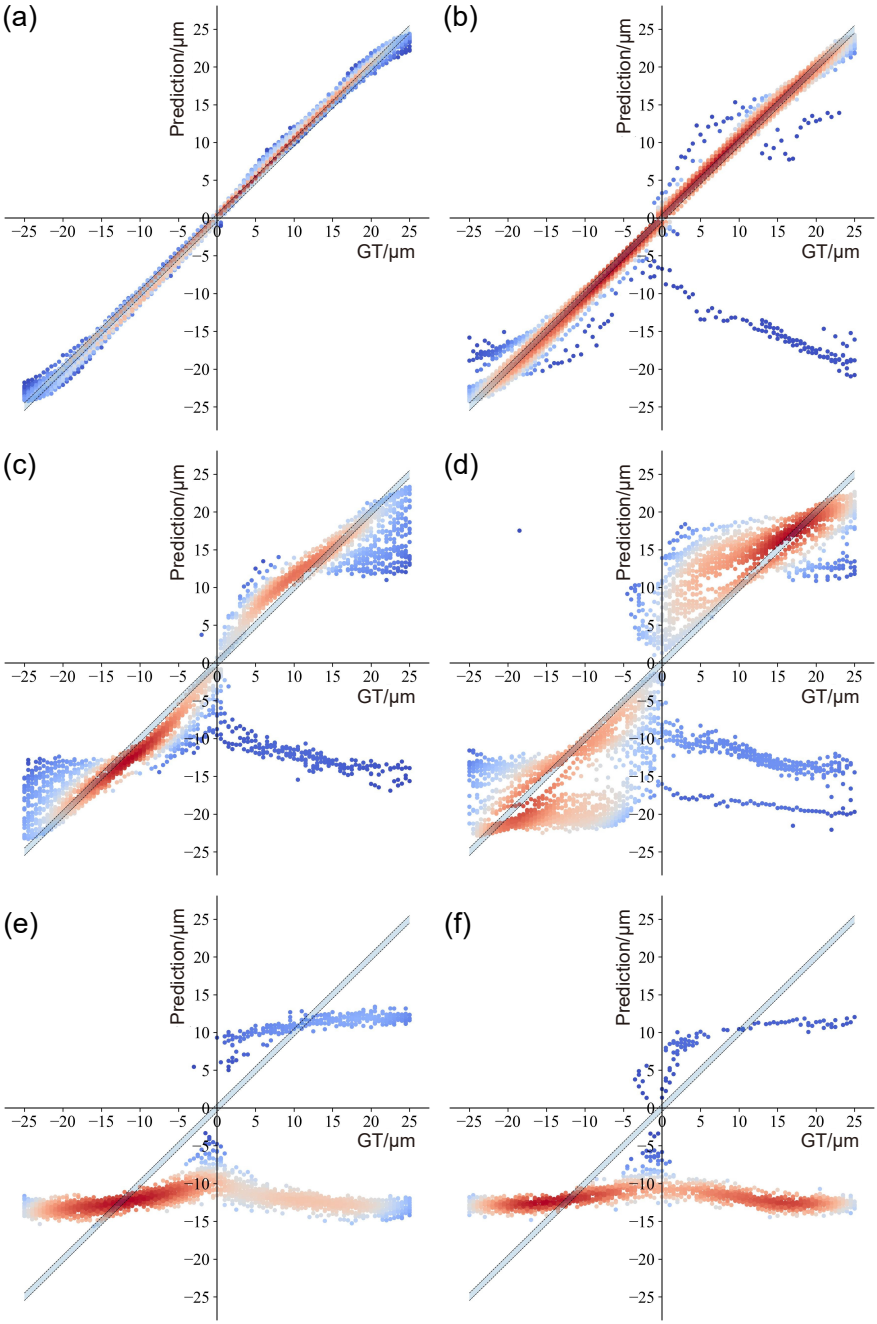

Fig. S17 Regression plots of Li et al. [4]. (a) shows the results on cell samples in dense scenarios, (b) shows the results on tissue samples in dense scenarios, (c) shows the results on cell samples in sparse scenarios, (d) shows the results on tissue samples in sparse scenarios, (e) shows the results on cell samples in extremely sparse scenarios, (f) shows the results on tissue samples in extremely sparse scenarios.

In Fig. S13 to S17, our method, SparseFocus, demonstrates superior performance across all levels of sparsity, irrespective of whether the sample is a cell or tissue. While other baseline methods maintain reasonable accuracy under dense conditions, they exhibit a marked increase in prediction errors as sparsity increases. In scenarios of extreme sparsity, these baseline methods become largely ineffective.

S8 Ablation Experiments on DPN

We conducted ablation experiments on the DPN by reducing the depth of its feature extraction layers from the original configuration of (3, 3, 9, 3) to two shallower architectures: (2, 2, 6, 2) and (1, 1, 3, 1). These models are denoted as DPN_{full}, DPN_{2262}, and DPN_{1131}, respectively. The models were trained and evaluated on the same dataset. To visually compare their performance, we present defocus prediction regression plots for the three models in Fig. S18. It can be observed that as the model depth decreases, the number of outliers increases across all sparsity scenarios. This degradation is relatively moderate under dense sampling conditions (Fig. S18(a) to S18(c)), but becomes more evident in sparse (Fig. S18(d) to S18(f)) and extremely sparse scenarios (Fig. S18(g) to S18(i)).

We further evaluated the three models using the three metrics introduced in the main text, as summarized in Tables S2 to S4. The number of parameters and computational cost for each model are also provided in Table S5. The results indicate that both DPN_{1131} and DPN_{2262} exhibit performance degradation compared to DPN_{full}. In particular, under extremely sparse scenarios, the MAE and standard deviation increase noticeably as the network depth decreases, rising from 0.60 ± 0.46 for DPN_{full}, to 0.85 ± 1.84 for DPN_{2262}, and further to 0.92 ± 2.12 for DPN_{1131}. We note that in terms of prediction accuracy (reflected by MAE and standard deviation), DPN_{1131} outperforms DPN_{2262} on cell samples with both dense and sparse content. Regression plot analysis indicates that this behavior stems from incorrect defocus direction predictions in a small number of cases with large defocus distances.

Table S2 Performance comparison of defocus distance estimation across different sparsity levels using *MAE*.

MAE (µm)	Dense	Sparse	Ex-sparse
Cell			
DPN_{1131}	0.40 ± 0.43	0.74 ± 1.43	0.92 ± 2.12
DPN_{2262}	0.50 ± 0.40	0.80 ± 1.89	0.85 ± 1.84
DPN_{full}	0.37 ± 0.29	0.51 ± 0.40	0.60 ± 0.46
Tissue			
DPN_{1131}	0.83 ± 1.52	0.85 ± 1.08	0.68 ± 2.08
DPN_{2262}	0.81 ± 1.58	0.88 ± 1.09	0.60 ± 1.02
DPN_{full}	0.49 ± 0.40	0.43 ± 0.33	0.45 ± 0.36

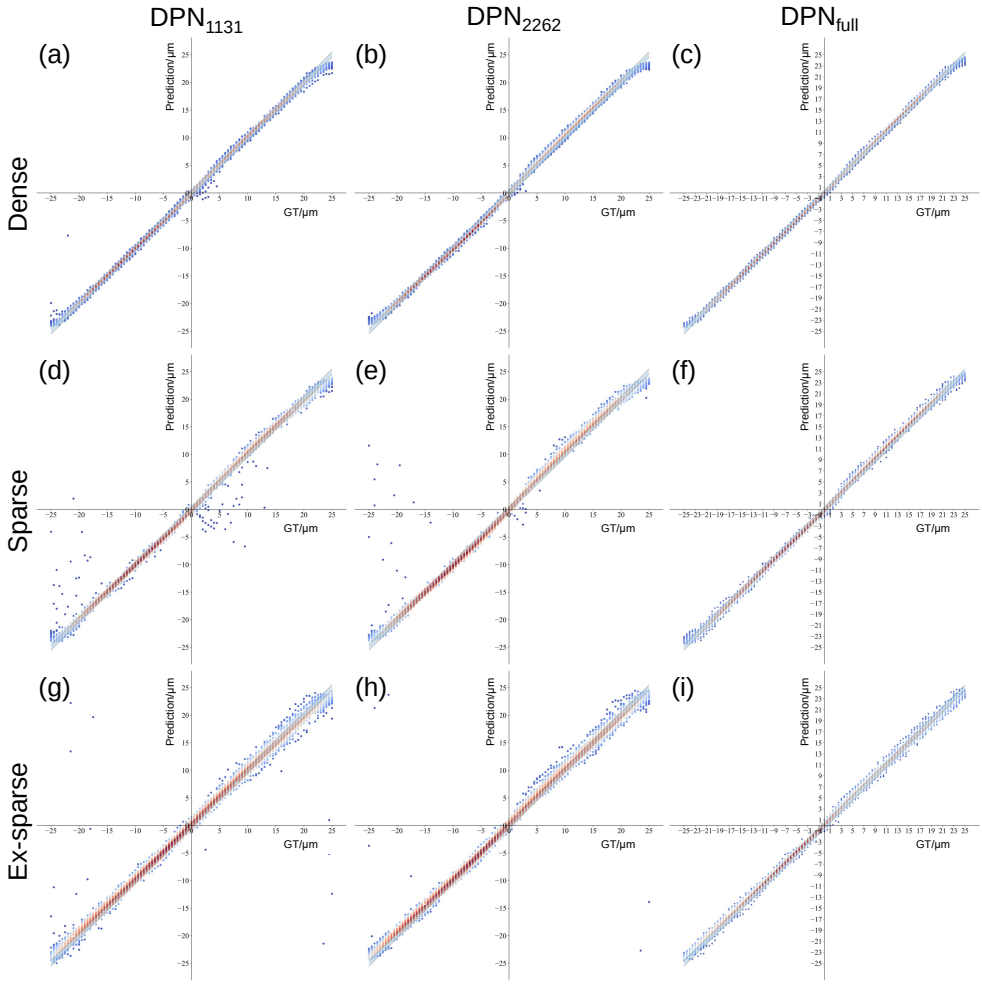

Fig. S18 Regression plots of DPN_{1131}, DPN_{2262} and DPN_{full}. (a) shows the results of DPN_{1131} in dense scenarios, (b) shows the results of DPN_{2262} in dense scenarios, (c) shows the results of DPN_{full} in dense scenarios, (d) shows the results of DPN_{1131} in sparse scenarios, (e) shows the results of DPN_{2262} in sparse scenarios, (f) shows the results of DPN_{full} in sparse scenarios, (g) shows the results of DPN_{1131} in extremely sparse scenarios, (h) shows the results of DPN_{2262} in extremely sparse scenarios, (i) shows the results of DPN_{full} in extremely sparse scenarios.

Table S3 Performance comparison of *DoF-Acc* across different sparsity levels for cell and tissue samples.

DoF-Acc (%)	Dense			Sparse			Ex-sparse		
	1/3	1/2	1	1/3	1/2	1	1/3	1/2	1
				Cell					
DPN_{1131}	35.24	50.98	79.49	25.80	36.60	62.29	17.03	25.73	47.38
DPN_{2262}	24.86	37.68	68.64	22.64	35.55	55.14	18.66	27.31	48.95
DPN_{full}	35.99	50.81	80.46	26.58	37.60	66.55	23.18	33.73	58.23
				Tissue					
DPN_{1131}	18.66	27.51	51.55	15.96	23.41	48.00	23.39	35.71	63.73
DPN_{2262}	17.43	27.33	51.55	15.47	23.49	46.58	23.56	34.29	60.96
DPN_{full}	26.54	39.38	69.72	27.53	40.85	74.69	28.16	41.74	71.71

Table S4 Performance comparison using *DSS* across different sparsity levels for cell and tissue samples.

DSS (%)	Cell			Tissue		
	Dense	Sparse	Ex-sparse	Dense	Sparse	Ex-sparse
DPN_{1131}	99.64	98.32	99.03	98.54	98.18	98.53
DPN_{2262}	99.73	99.27	99.36	98.69	98.34	98.87
DPN_{full}	100.00	99.80	99.96	99.77	99.91	99.94

Table S5 The number of parameters and computational cost for three models.

	DPN_{1131}	DPN_{2262}	DPN_{full}
Params	19.71M	35.22M	50.72M
MACs	392.77M	445.79M	518.81M

References

[1] Dastidar, T.R., Ethirajan, R.: Whole slide imaging system using deep learning-based automated focusing. Biomedical Optics Express **11**(1), 480–491 (2020)

[2] Liao, J., Chen, X., Ding, G., Dong, P., Ye, H., Wang, H., Zhang, Y., Yao, J.: Deep learning-based single-shot autofocus method for digital microscopy. Biomedical Optics Express **13**(1), 314–327 (2022)

[3] Jiang, S., Liao, J., Bian, Z., Guo, K., Zhang, Y., Zheng, G.: Transform-and multi-domain deep learning for single-frame rapid autofocusing in whole slide imaging. Biomedical optics express **9**(4), 1601–1612 (2018)

[4] Li, Q., Liu, X., Han, K., Guo, C., Jiang, J., Ji, X., Wu, X.: Learning to autofocus in whole slide imaging via physics-guided deep cascade networks. Optics Express **30**(9), 14319–14340 (2022)