Supplementary Information

A foundation model for multi-task cross-distribution restoration of fluorescence microscopy image

Qiqi Lu[1,2,3], Xiuli Liu[4], Qianjin Feng[1,2,3], Shaoqun Zeng[4,*] and Shenghua Cheng[1,2,3,*]

[1]School of Biomedical Engineering, Southern Medical University, Guangzhou, China.

[2]Guangdong Provincial Key Laboratory of Medical Image Processing, Southern Medical University, Guangzhou, China.

[3]Guangdong Province Engineering Laboratory for Medical Imaging and Diagnostic Technology, Southern Medical University, Guangzhou, China.

[4]MOE Key Laboratory for Biomedical Photonics, Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan, China.

*Correspondence: sqzeng@mail.hust.edu.cn, chengsh2023@smu.edu.cn

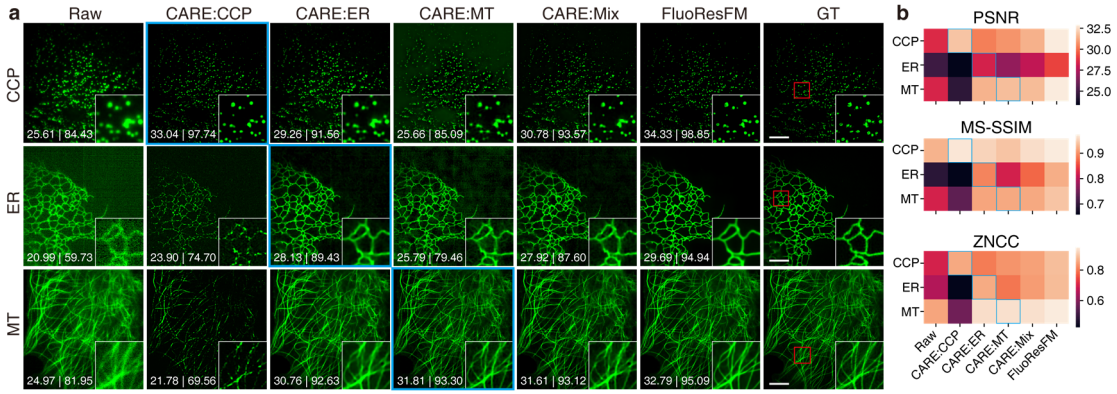
Supplementary Notes

Supplementary Note 1: Structure type prediction based on image retrieval

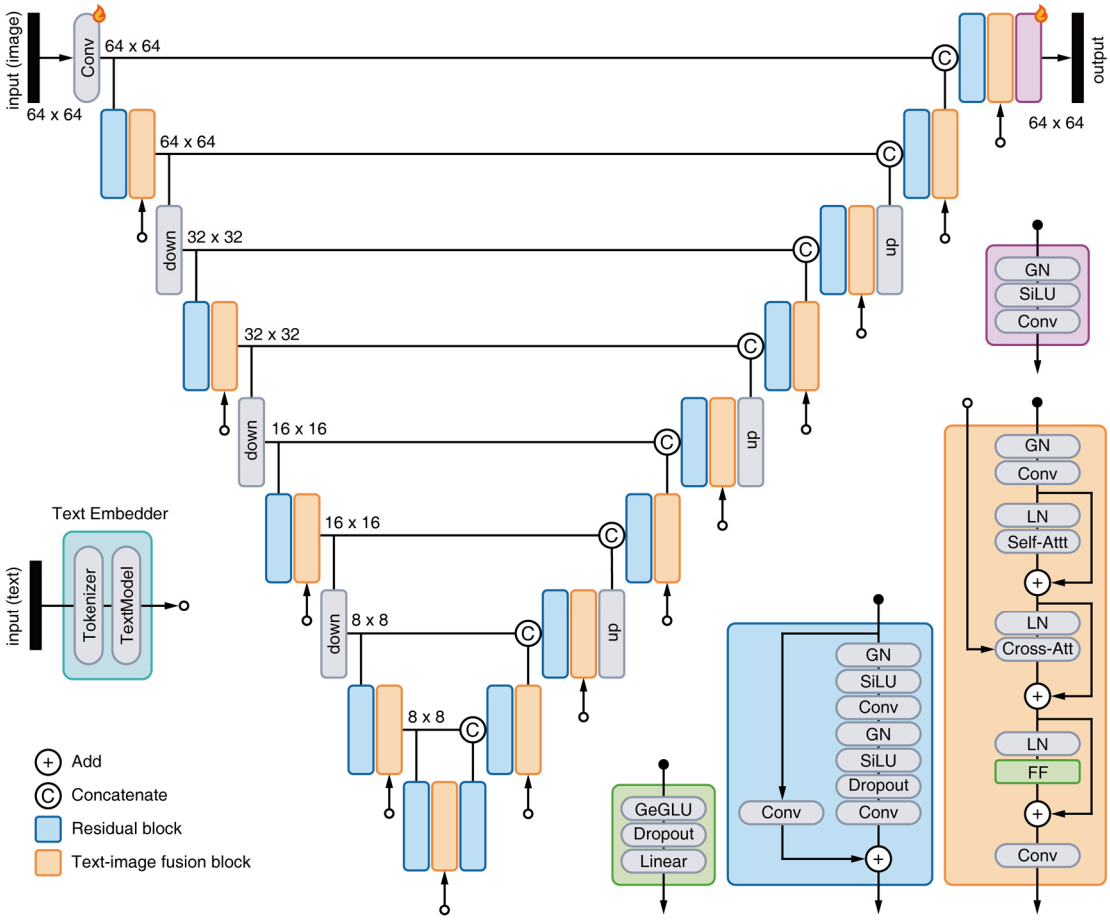
Given the critical role of the structure type prompt in restoration, employing a structure type classifier to help general users to recognize the structure type in the input image would be highly beneficial. We adopted an image retrieval-based strategy for structure type classification (Supplementary Fig. 7a). Image retrieval aims to find images in a database that are most similar to a given query image[1]. To assess the effectiveness of this approach, we constructed an image database of 52,544 images (each with a size of 224×224 pixels) from the training datasets (Supplementary Fig. 7b) and used 7,708 images from the test datasets (Supplementary Fig. 7c) to evaluate its classification accuracy. We used the pretrained image encoder from BiomedCLIP to precompute embeddings for all the images in the database[2]. The resulting high-level embeddings are highly discriminative, with images of the same structure type forming compact clusters in the embedding space (Supplementary Fig. 7d).

For a given query image, we used the same image encoder to compute its embedding and calculated its cosine similarities with all the image embeddings in the database. We then retrieved the k most similar images, and assigned the structure type that appeared most frequently among these candidates as the predicted label of the query image (Supplementary Fig. 7a). This approach achieved a high accuracy of 91.46% on the test set (Supplementary Fig. 7e). Most classification errors arose from the low-SNR of images or the high structural similarity between different structures (for example, CCP, NPC, and myosin-IIA all exhibit a spot-like structure). To assess the benefit of this classifier for non-expert users, 1,800 images randomly collected from the test set were manually annotated by a general user. Although this user had been trained for a short period to distinguish different types of structures, the obtained accuracy of 71.44% was significantly lower than the image retrieval-based classification accuracy of 91.83%. (Supplementary Fig. 7f). These results indicated that the proposed classifier could effectively reduce prompt-image mismatches compared to manual specification of structure type by general users.

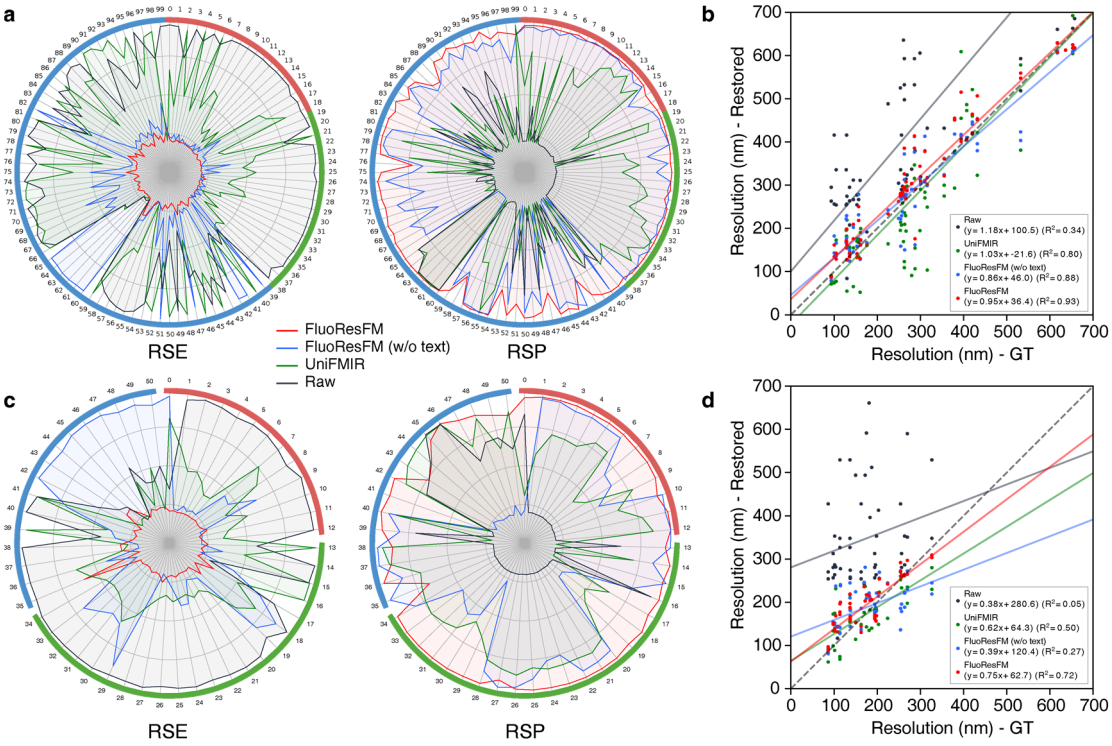
Supplementary Figures



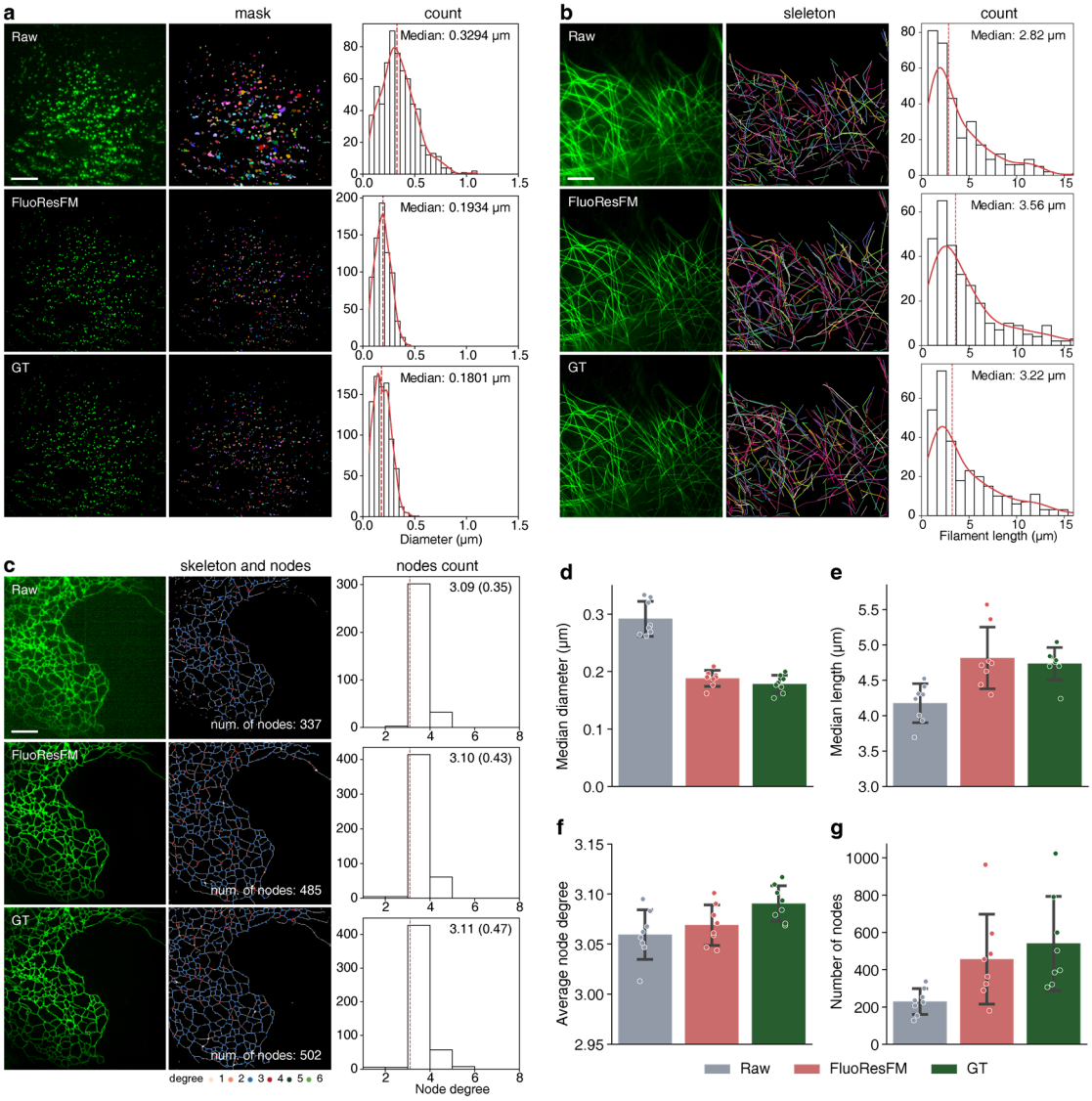
Supplementary Fig. 1 | Influence of data distribution bias on network performance. a, Effect of data distribution bias between training and testing stages on the super-resolution restoration performance of the structure-specific deep learning model CARE. CARE was trained on datasets containing CCP (CARE:CCP), ER (CARE:ER), MT (CARE:MT), or a mixed set of these structures (CARE:Mix), and tested on datasets with CCP, ER, and MT structures. PSNR and MS-SSIM values are shown in the left bottom corner of each image. Raw, input WF image. GT, ground truth SIM image. The magnified region (red box) is shown in the right bottom corner. **b**. Comparison of different methods in terms of average PSNR, MS-SSIM, and ZNCC on datasets with different structures (each with n = 8 test images). The blue boxes highlight the results obtained when the training and testing data have the same distribution. Scale bar: 5 μm (**a**). Source data are provided as a Source Data file.



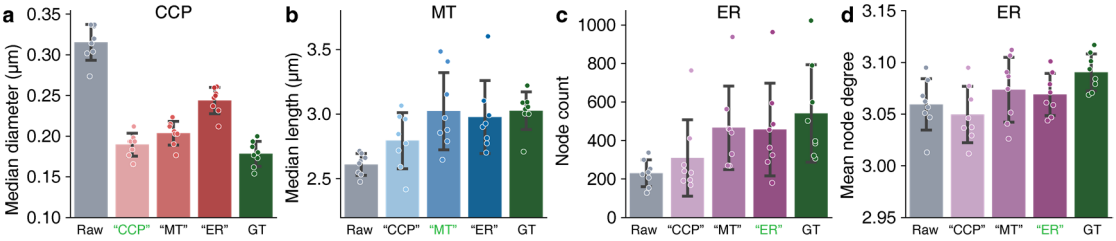
Supplementary Fig. 2 | Detailed architecture of FluoResFM. FluoResFM takes the low-quality image and its corresponding textual prior information as inputs and outputs the restored high-quality image. It employs a text-conditioned U-Net as backbone, which consists of an encoder, a decoder, and skip connections. The encoder begins with a convolution layer, and the decoder ends with a convolution layer to generate the restored images. Each scale of the encoder and decoder contains a residual block and a text-image fusion block. The text prompt is projected into a text embedding using a text embedder and then injected into the cross-attention layer within the text-image fusion block. Only blocks labeled with “fire” markers are fine-tuned during the fine-tuning stage.



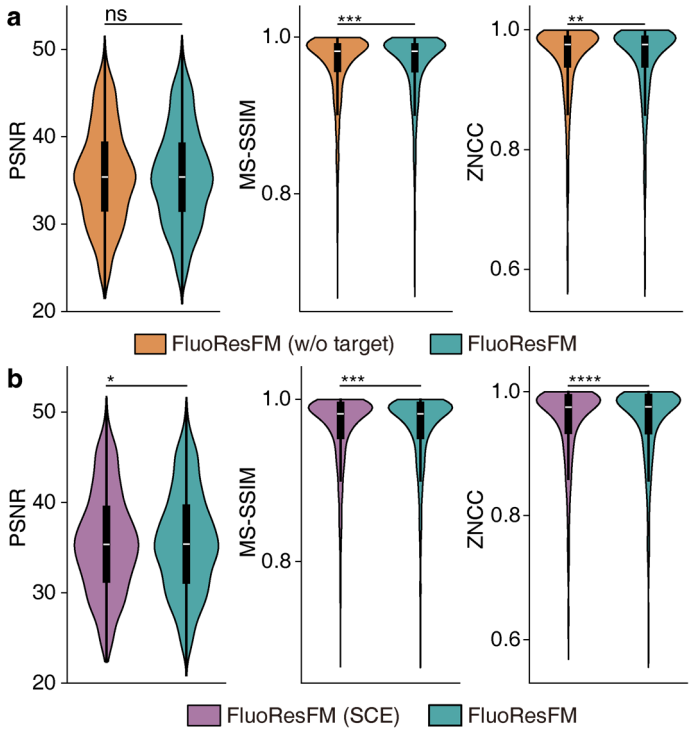
Supplementary Fig. 3 | Performance comparison on internal and external datasets. a, Quantitative performance comparison of different methods on internal datasets across three tasks (SR: 0-18, DCV: 19-38, DN: 39-99) in terms of average RSE and RSP. Only 100 datasets are shown due to the space constraints. The metric values are normalized based on the minimum and maximum values of each dataset (Methods). **b**, Consistency between the resolution of restored images and ground truth reference images on internal datasets. Each point represents the average resolution across a dataset. Datasets with resolution estimates far worse than the diffraction-limited widefield resolution were excluded as outliers. The gray dashed line represents the identity line, and the solid line represents the regression line. The fitted linear regression model and coefficient of determination (R^2) value are shown in the legend. **c**, Quantitative performance comparison of different methods on external unseen datasets across three tasks (SR: 0-12, DCV: 13-34, DN: 35-50) in terms of average RSE and RSP. **d**, Consistency between the resolution of restored images and ground truth reference images on external datasets. Source data are provided as a Source Data file.



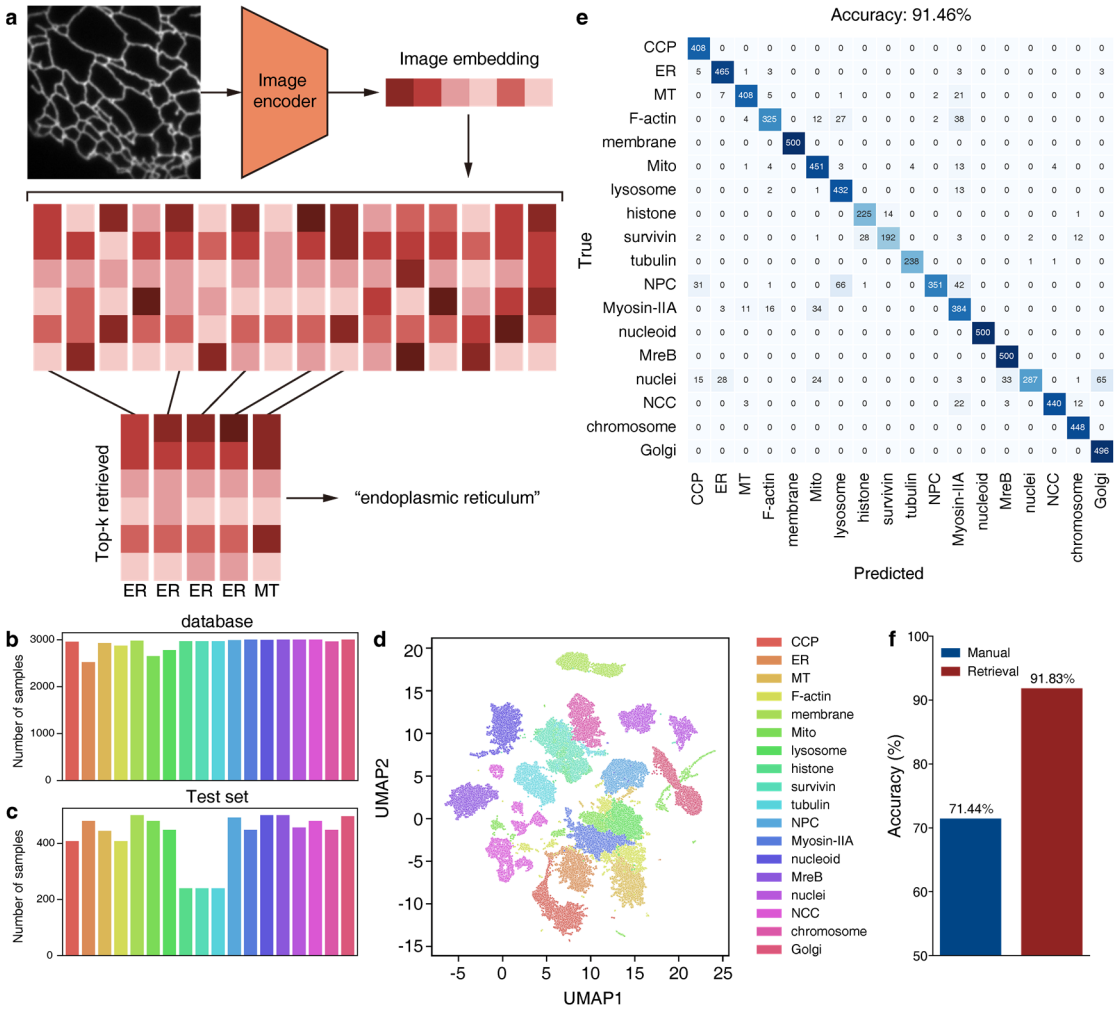
Supplementary Fig. 4 | Evaluation of the utility of FluoResFM for measuring CCP diameter, filament length in MT network, and node degree/count in ER network. a, Segmentation of CCPs and their diameter distribution. The red line in the histogram represents the kernel density estimate of the distribution, and the dashed line indicates the median diameter, the value of which is labeled in the top-right corner. **b**, Tacking of filaments in MT network and their length distribution. The dashed line in the histogram represents the median filament length, the value of which is labeled in the top-right corner. **c**, Detection of ER nodes and their node degree distribution. The mean and standard deviation of node degree are shown in the top-right corner of the histogram. **d**,**e**,**f**,**g**, Comparison of median diameter of CCPs (**d**), median filament length in MT network (**e**), average node degree (**f**), and number of detected nodes in ER network (**g**) measured from raw input, restored, and GT images. The error bar represents the mean value ± standard deviation of the metric values (n = 8 images). Each point represents the value for a sample. Scale bar: 5 μm. Source data are provided as a Source Data file.



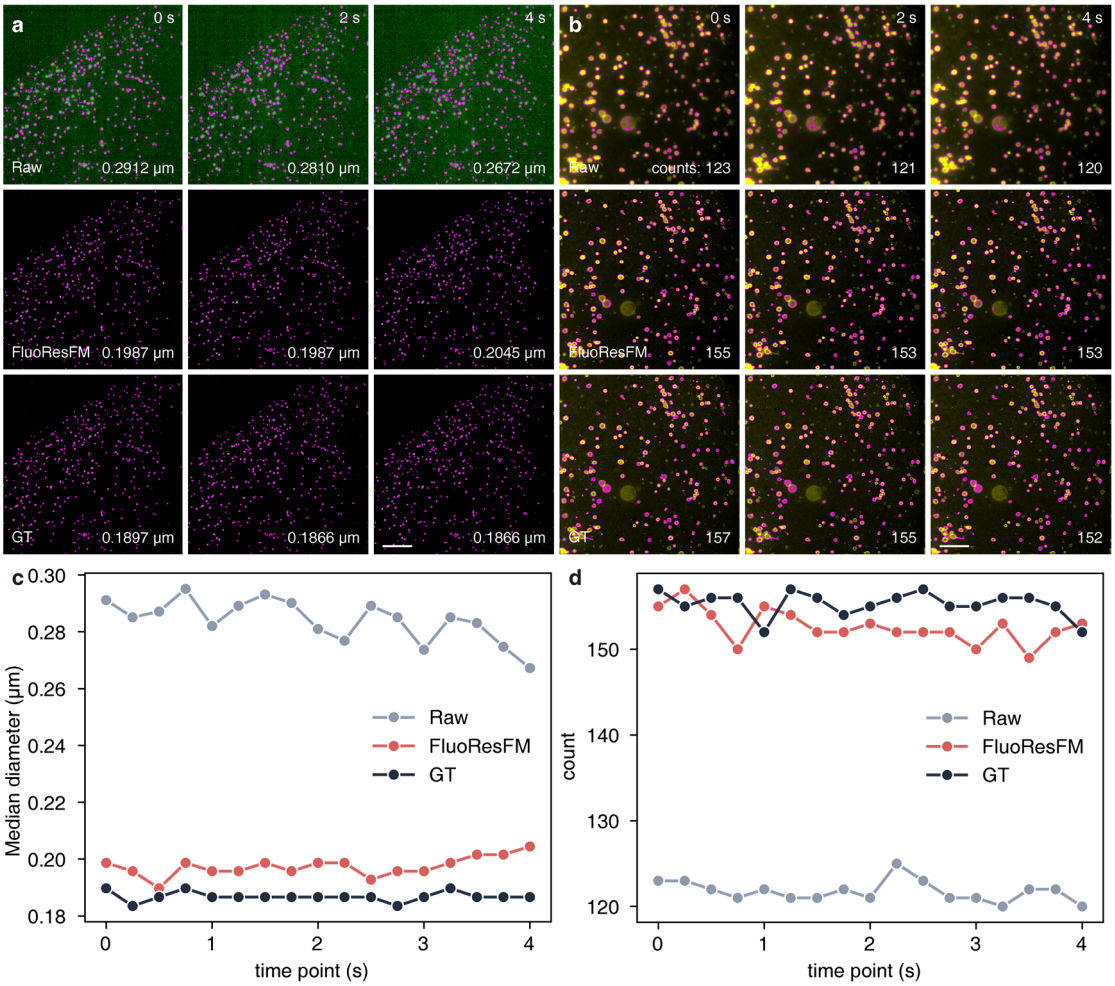
Supplementary Fig. 5 | Quantitative evaluation of the influence of input structure type prompt mismatch. a, Median CCP diameter measured from images restored via FluoResFM with different input structure prompts (i.e., "CCP", "MT", and "ER") (n = 8 images). **b**, Median filament length of MT network (n = 8 images). **c**,**d**, Node count (**c**) and mean node degree (**d**) of ER network (n = 8 images). The matched prompts are marked in green. Error bar represents the mean value ± standard deviation of the metric values. Each point corresponds to a single sample. Source data are provided as a Source Data file.



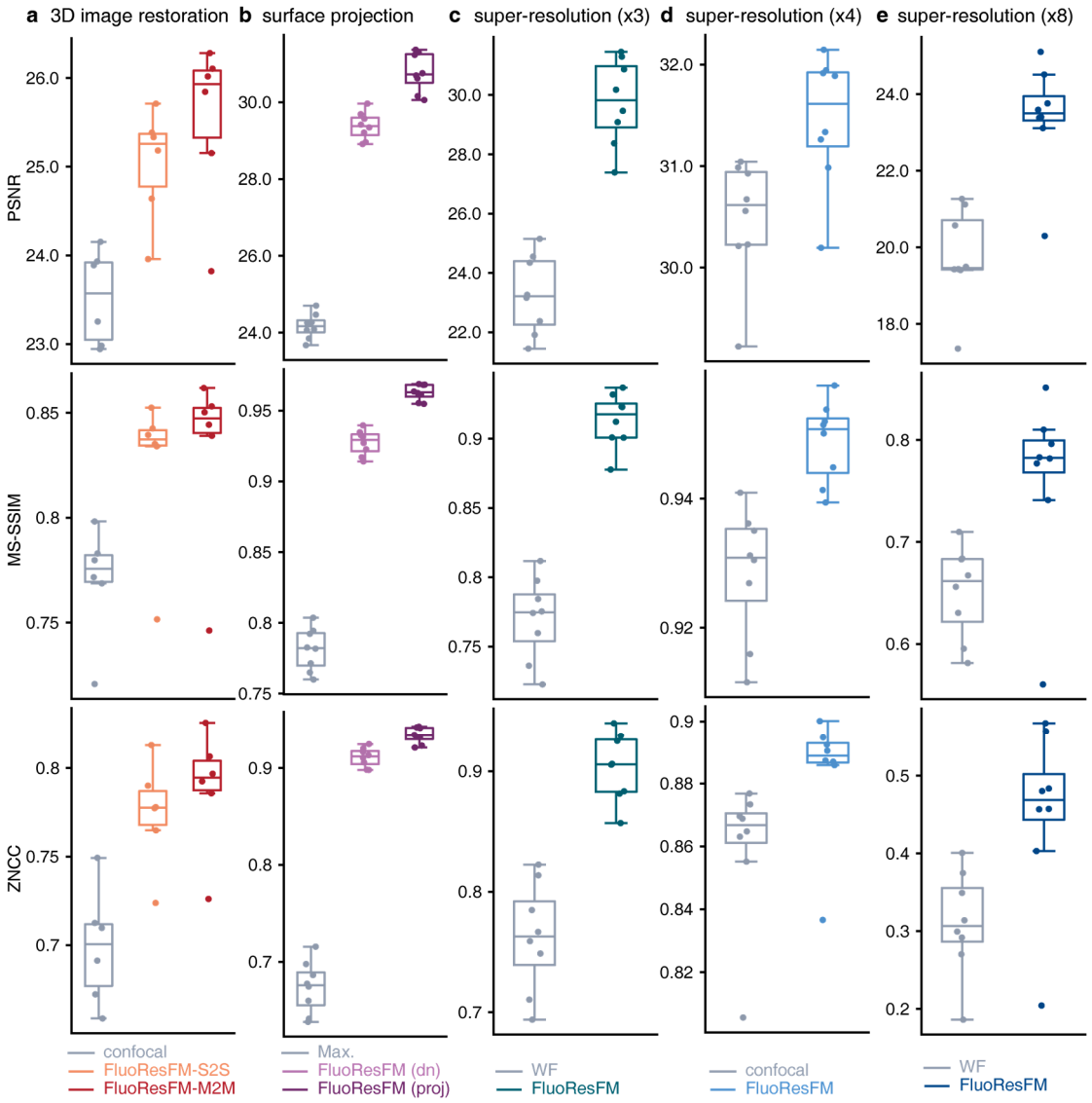
Supplementary Fig. 6 | Performance comparison between FluoResFM with its variants on internal datasets. a, Comparison to FluoResFM using prompts without target metadata (FluoResFM (w/o target)) in terms of PSNR, MS-SSIM, and ZNCC. **b**, Comparison to FluoResFM with structured categorical embeddings (FluoResFM (SCE)). The box extends from the first quartile to the third quartile of the data, with a white line at the median. The whiskers extend from the box to the farthest data point lying within 1.5× the inter-quartile range. Statistical significance is determined by Wilcoxon signed-rank test (n = 2,387) (****$p < 0.0001$, ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$). Source data are provided as a Source Data file.



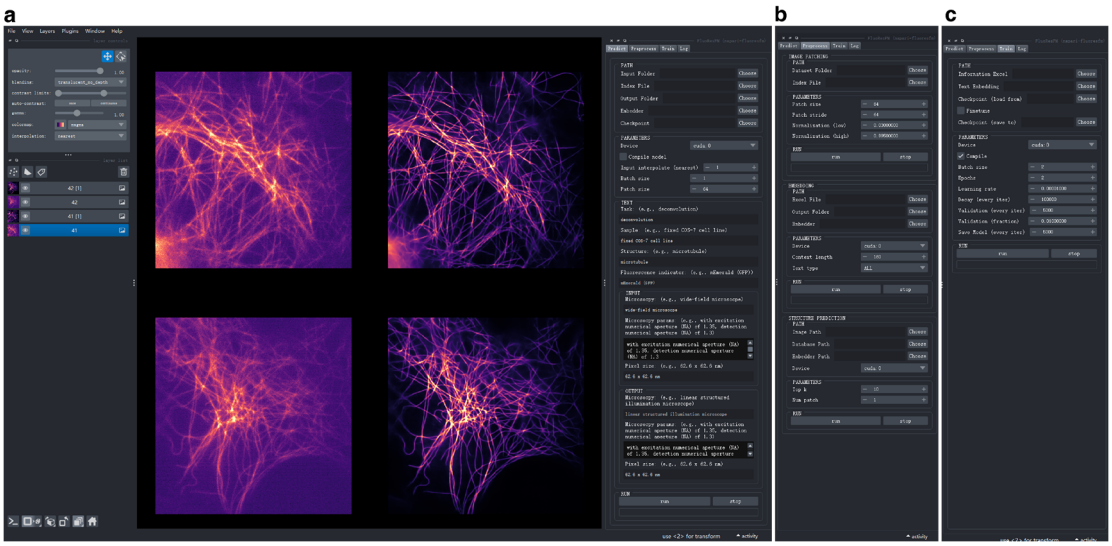
Supplementary Fig. 7 | Structure type prediction based on image retrieval. a, Schematic for image-to-image retrieval. The query image is first embedded using a pretrained image encoder from BiomedCLIP, then similarity in the embedding space is computed between the query image and all the samples in the database. The top-*k* most similar images are retrieved. The most frequent class among them is treated as the label of the query image. **b**,**c**, Number of samples for different types of imaging structures in the database (52,544 images in total) (**b**) and in the test set (7,708 images in total) (**c**). **d**, Uniform manifold approximation and projection (UMAP) visualization of image embeddings in the database. **e**, Confusion matrix on the test set (n = 7,708 images). **f**, Accuracy comparison between manual annotation and classification based on image retrieval (n = 1,800 images). Source data are provided as a Source Data file.



Supplementary Fig. 8 | Evaluation of CCP diameter and lysosome count across different time points. a, Segmentation of CCPs in raw input, restored, and ground truth (GT) images. The time point is labeled in the top-right corner of raw image. The median diameter of CCPs is shown in the bottom-right corner of each image. **b**, Segmentation of lysosomes in raw input, restored, and GT images. The count of lysosomes is shown in the bottom-right corner of each image. **c**, Median CCP diameter across all the time points. **d**, Lysosome count across all the time points. Scale bar: 5 μm. Source data are provided as a Source Data file.



Supplementary Fig. 9 | Quantitative evaluation of FluoResFM when transferred to other restoration tasks via fine-tuning on a single sample. a, Comparison of the restoration performance between different strategies used in 3D image restoration in terms of PSNR, MS-SSIM, and ZNCC (n = 6 images). FluoResFM-S2S restores the low-quality 3D volume slice-by-slice, while FluoResFM-M2M treats the 3D volume input as a multi-channel 2D input and simultaneously restored all the image slices. **b**, Comparison of the performance between different strategies used in surface projection task. Max., maximum intensity projection (n = 8 images). FluoResFM (dn) first projects the low-SNR input volume to a low-SNR 2D surface image via maximum intensity projection and then denoises the 2D surface image using FluoResFM. FluoResFM (proj) treats the 3D volume input as a multi-channel 2D input and directly projects it into a 2D surface image using FluoResFM. **c**,**d**,**e**, Comparison of image quality before and after restoration using FluoResFM in the super-resolution tasks with scale factors of 3× (**c**), 4× (**d**), and 8× (**e**) (n = 8 image in each dataset). The box extends from the first quartile to the third quartile of the data, with a white line at the median. The whiskers extend from the box to the farthest data point lying within 1.5× the inter-quartile range. Source data are provided as a Source Data file.



Supplementary Fig. 10 | Interactive interface of the napari plugin for FluoResFM. The developed napari plugin can be used to implement data preprocessing, model training, and model prediction. **a**, The widget used for restoring low-quality images using a pre-trained FluoResFM model. Two pairs of representative low-quality input and high-quality restored images are shown in the viewer. **b**, The widget used for data preprocessing, which can perform image patching, text embedding, and structure type prediction. **c**, The widget used for model training and fine-tuning.

Supplementary Tables

Supplementary Table 1 | Overview of the training/internal dataset details. The image restoration task, imaging object, imaging device (microscopy), and the download link of each dataset are provided. DN, denoising; DCV, deconvolution; SR, super-resolution; CCP, clathrin-coated pit; ER, endoplasmic reticulum; MT, microtubule; Mito, mitochondria; NCC, neural crest cell; NPC, nuclear pore complex; BPAE, bovine pulmonary artery endothelial cell; WF, wide-field; SIM, structured illumination microscopy; STED, stimulated emission depletion microscopy.

Dataset	Task	Imaging object	Microscopy	Link
BioSR	DN, DCV, SR	CCP	WF, SIM	https://figshare.com/articles/dataset/BioSR/13264793
		ER	WF, SIM	https://figshare.com/articles/dataset/BioSR/13264793
		F-actin	WF, SIM	https://figshare.com/articles/dataset/BioSR/13264793
		MT	WF, SIM	https://figshare.com/articles/dataset/BioSR/13264793
BioSR+	DN	CCP	WF	https://zenodo.org/records/7115540
		ER	WF	https://zenodo.org/records/7115540
		F-actin	WF	https://zenodo.org/records/7115540
		MT	WF	https://zenodo.org/records/7115540
		myosin-IIA	WF	https://zenodo.org/records/7115540
CARE	DN	nuclei (planaria)	confocal	https://doi.org/10.17617/3.FDFZOF
		nuclei (Tribolium)	multi-photon	https://doi.org/10.17617/3.FDFZOF
	SR	histone (Drosophila)	light-sheet	https://doi.org/10.17617/3.FDFZOF
		nuclei (retina)	confocal	https://doi.org/10.17617/3.FDFZOF
		nuclear-envelope (retina)	confocal	https://doi.org/10.17617/3.FDFZOF
		nuclei/membrane	two-photon	https://doi.org/10.17617/3.FDFZOF
	DCV	granule	WF	https://doi.org/10.17617/3.FDFZOF
		tubulin	WF	https://doi.org/10.17617/3.FDFZOF
		tubulin	confocal	https://doi.org/10.17617/3.FDFZOF
DeepBacs	DN	nucleoid	confocal	https://github.com/HenriquesLab/DeepBacs/wiki
		MreB	confocal	https://github.com/HenriquesLab/DeepBacs/wiki
	DCV, SR	membrane (E. coli)	WF, SIM	https://github.com/HenriquesLab/DeepBacs/wiki
		membrane (S. aureus)	WF, SIM	https://github.com/HenriquesLab/DeepBacs/wiki
W2S	DN, DCV, SR	Mito	WF, SIM	https://zenodo.org/records/3895807
		lysosome	WF, SIM	https://zenodo.org/records/3895807
		F-actin	WF, SIM	https://zenodo.org/records/3895807
SR-CACO-2	DN, SR	histone	confocal	https://github.com/sbelharbi/sr-caco-2
		survivin	confocal	https://github.com/sbelharbi/sr-caco-2
		tubulin	confocal	https://github.com/sbelharbi/sr-caco-2
FMD	DN	nuclei (BPAE)	confocal	https://github.com/yinhaoz/denoising-fluorescence
		F-actin (BPAE)	confocal	https://github.com/yinhaoz/denoising-fluorescence
		Mito (BPAE)	confocal	https://github.com/yinhaoz/denoising-fluorescence
		NCC (zebrafish)	confocal	https://github.com/yinhaoz/denoising-fluorescence
		chromosome (mice)	confocal	https://github.com/yinhaoz/denoising-fluorescence
		chromosome (mice)	two-photon	https://github.com/yinhaoz/denoising-fluorescence
		nuclei (BPAE)	two-photon	https://github.com/yinhaoz/denoising-fluorescence
		F-actin (BPAE)	two-photon	https://github.com/yinhaoz/denoising-fluorescence
		Mito (BPAE)	two-photon	https://github.com/yinhaoz/denoising-fluorescence
		nuclei (BPAE)	WF	https://github.com/yinhaoz/denoising-fluorescence
		F-actin (BPAE)	WF	https://github.com/yinhaoz/denoising-fluorescence
		Mito (BPAE)	WF	https://github.com/yinhaoz/denoising-fluorescence
VMSIM5	DCV, SR	Mito	WF, SIM	https://zenodo.org/records/3295829
RCAN-in	DCV, SR	MT	confocal, STED	https://zenodo.org/records/4624364#.YF4lBa9Kgal
		NPC	Confocal, STED	https://zenodo.org/records/4624364#.YF4lBa9Kgal
	DN	golgi	SIM	https://zenodo.org/records/4624364#.YF4lBa9Kgal

Supplementary Table 2 | Overview of the external evaluation datasets. DN, denoising; DCV, deconvolution; SR, super-resolution; MT, microtubule; Mito, mitochondria; ER, endoplasmic reticulum; CCP, clathrin-coated pit; WF, wide-field; SIM, structured illumination microscopy; iSIM, instant structured illumination microscopy.

Dataset	Task	Imaging object	Microscopy	Link
SIM	DCV	F-actin	WF, SIM	http://www.cellimagelibrary.org/images/7052
		MT	WF, SIM	http://www.cellimagelibrary.org/images/36797
VMSIM3	DCV, SR	Mito	WF, SIM	https://zenodo.org/records/3295829
VMSIM488	DCV	microsphere	WF, SIM	https://zenodo.org/records/3295829
VMSIM568	DCV	microsphere	WF, SIM	https://zenodo.org/records/3295829
VMSIM647	DCV	microsphere	WF, SIM	https://zenodo.org/records/3295829
RCAN-ex	DN	F-actin	iSIM	https://zenodo.org/records/4624364#.YF4lBa9Kgal
		ER	iSIM	https://zenodo.org/records/4624364#.YF4lBa9Kgal
		lysosome	iSIM	https://zenodo.org/records/4624364#.YF4lBa9Kgal
		Mito	iSIM	https://zenodo.org/records/4624364#.YF4lBa9Kgal
		MT	iSIM	https://zenodo.org/records/4624364#.YF4lBa9Kgal
BioTISR	DN, DCV, SR	CCP	WF, SIM	https://zenodo.org/records/14760518
		F-actin	WF, SIM	https://zenodo.org/records/14760518
		lysosome	WF, SIM	https://zenodo.org/records/14760518
		MT	WF, SIM	https://zenodo.org/records/14760518
		Mito	WF, SIM	https://zenodo.org/records/14760518

Supplementary Table 3 | Overview of the datasets used for 3D image restoration, surface projection, isotropic reconstruction, and super-resolution tasks with various scale factors. DCV, deconvolution; Proj, surface projection; ISO, isotropic reconstruction; SR, super-resolution; MT, microtubule; WF, wide-field; SIM, structured illumination microscopy; STED, stimulated emission depletion microscopy.

Dataset	Task	Imaging object	Microscopy	Link
RCAN-3D	3D DCV	MT	confocal, STED	https://zenodo.org/records/4624364#.YF4lBa9Kgal
CARE-proj	Proj	membrane	confocal	https://doi.org/10.17617/3.FDFZOF
CARE-iso	ISO	histone	light-sheet	https://doi.org/10.17617/3.FDFZOF
BioSR-3x	SR (×3)	F-actin	WF, SIM	https://figshare.com/articles/dataset/BioSR/13264793
TA-GAN	SR (×4)	synaptic protein	confocal, STED	https://s3.valeria.science/flclab-tagan/index.html
DL-SMLM	SR (×8)	MT	WF, STORM	https://doi.org/10.6084/m9.figshare.26879218.v1

Supplementary Table 4 | Overview of datasets used for segmentation performance evaluation. The "task" refers to the applied restoration task before segmentation using Nellie or Cellpose-SAM. DN, denoising; DCV, deconvolution; SR, super-resolution; CCP, clathrin-coated pit; Mito, mitochondria; BPAE, bovine pulmonary artery endothelial cell; WF, wide-field; SIM, structured illumination microscopy.

Dataset	Task	Imaging object	Microscopy	Link
CARE	DN	nuclei (planaria)	confocal	https://doi.org/10.17617/3.FDFZOF
		nuclei (Tribolium)	multi-photon	https://doi.org/10.17617/3.FDFZOF
	SR	nuclei/membrane	two-photon	https://doi.org/10.17617/3.FDFZOF
DeepBacs	DN	nucleoid	confocal	https://github.com/HenriquesLab/DeepBacs/wiki
	DCV	membrane (E. coli)	WF, SIM	https://github.com/HenriquesLab/DeepBacs/wiki
		membrane (S. aureus)	WF, SIM	https://github.com/HenriquesLab/DeepBacs/wiki
SR-CACO-2	DN, SR	histone	confocal	https://github.com/sbelharbi/sr-caco-2
FMD	DN	nuclei (BPAE)	confocal	https://github.com/yinhaoz/denoising-fluorescence
		F-actin (BPAE)	confocal	https://github.com/yinhaoz/denoising-fluorescence
		Mito (BPAE)	confocal	https://github.com/yinhaoz/denoising-fluorescence
BioTISR	SR	CCP	WF, SIM	https://zenodo.org/records/14760518
		F-actin	WF, SIM	https://zenodo.org/records/14760518
		lysosome	WF, SIM	https://zenodo.org/records/14760518
HL60	DN, DCV	nuclei	confocal	https://bbbc.broadinstitute.org/BBBC024
Scaffold-A549	DN	nuclei	confocal	https://github.com/Kaiseem/Scaffold-A549
GranuSEG	DN	nuclei	confocal	https://cbia.fi.muni.cz/datasets/
ColonTissue	DCV	nuclei	confocal	https://cbia.fi.muni.cz/datasets/

Supplementary Table 5 | Detail settings of FluoResFM and comparison models during training and fine-tuning. As UniFMIR model can only be trained with a batch size of 1, it used more training iterations to go through all the training patches.

	Model	Optimizer	Initial learning rate	Patch size	Batch size	Total training iterations
Fig. 1-6	FluoResFM	AdamW	1×10^{-5}	64×64	16	700,000
Fig. 2,3,5	UniFMIR	AdamW	1×10^{-4}	64×64	1	4,300,000
Extended Data Fig. 1	CARE:CCP	AdamW	1×10^{-4}	64×64	16	700,000
Extended Data Fig. 1	CARE:ER	AdamW	1×10^{-4}	64×64	16	700,000
Extended Data Fig. 1	CARE:MT	AdamW	1×10^{-4}	64×64	16	700,000
Extended Data Fig. 1	CARE:Mix	AdamW	1×10^{-4}	64×64	16	700,000
Fig. 5	CARE	AdamW	1×10^{-4}	64×64	16	32,000
Fig. 5	DFCAN	AdamW	1×10^{-4}	64×64	16	32,000
Fig. 5	FluoResFM (ft)	AdamW	1×10^{-5}	64×64	16	32,000
Fig. 5	UniFMIR (ft)	AdamW	1×10^{-4}	64×64	1	36,600
Fig. 6c	FluoResFM-M2M	AdamW	1×10^{-4}	6×64×64	16	32,500
Fig. 6d	FluoResFM (proj)	AdamW	1×10^{-5}	50×64×64	16	32,500
Fig. 6d	FluoResFM (dn)	AdamW	1×10^{-5}	64×64	16	32,500
Fig. 6e	FluoResFM	AdamW	1×10^{-5}	64×64	16	35,000
Fig. 6h,j	FluoResFM	AdamW	1×10^{-3}	64×64	16	32,000
Fig. 6i	FluoResFM	AdamW	1×10^{-3}	64×64	16	32,000

Supplementary Table 6 | Detailed hyper-parameter settings of the network.

		Hyper-params	Num. of Params
FluoResFM		N_{cin} =1, N_{cout} =1, N_c=320, N_{head} = 8, D_{cond}=768	664,793,281
Conv2D (input)		in_channels = N_{cin}, out_channels = N_c, kernel_size = 3	3200
Residual block		channels = N_c, out_channels = N_c (identity skip)	1,845,120
Text-image fusion block		channels = N_c, n_heads = N_{head}, n_layers=1, d_cond = D_{cond}	2,546,240
down		channels = N_c	921,920
Residual block		channels = N_c, out_channels = 2× N_c (conv skip)	5,738,240
Text-image fusion block		channels = 2× N_c, n_heads = N_{head}, n_layers=1, d_cond = D_{cond}	9,188,480
down		channels = 2× N_c	3,687,040
Residual block		channels = 2× N_c, out_channels = 4× N_c (conv skip)	22,945,280
Text-image fusion block		channels = 4× N_c, n_heads = N_{head}, n_layers=1, d_cond = D_{cond}	34,760,960
down		channels = 4× N_c	14,746,880
Residual block		channels = 4× N_c, out_channels = 4× N_c (identity skip)	29,498,880
Text-image fusion block		channels = 4× N_c, n_heads = N_{head}, n_layers=1, d_cond = D_{cond}	34,760,960
Residual block		channels = 4× N_c, out_channels = 4× N_c (identity skip)	29,498,880
Text-image fusion block		channels = 4× N_c, n_heads = N_{head}, n_layers=1, d_cond = D_{cond}	34,760,960
Residual block		channels = 4× N_c, out_channels = 4× N_c (identity skip)	29,498,880
Residual block		channels = 8× N_c, out_channels = 4× N_c (conv skip)	47,525,120
Text-image fusion block		channels = 4× N_c, n_heads = N_{head}, n_layers=1, d_cond = D_{cond}	34,760,960
Residual block		channels = 8× N_c, out_channels = 4× N_c (conv skip)	47,525,120
Text-image fusion block		channels = 4× N_c, n_heads = N_{head}, n_layers=1, d_cond = D_{cond}	34,760,960
up		channels = 4× N_c	14,746,880
Residual block		channels = 8× N_c, out_channels = 4× N_c (conv skip)	47,525,120
Text-image fusion block		channels = 4× N_c, n_heads = N_{head}, n_layers=1, d_cond = D_{cond}	34,760,960
Residual block		channels = 6× N_c, out_channels = 4× N_c (conv skip)	39,331,840
Text-image fusion block		channels = 4× N_c, n_heads = N_{head}, n_layers=1, d_cond = D_{cond}	34,760,960
up		channels = 4× N_c	14,746,880
Residual block		channels = 6× N_c, out_channels = 2× N_c (conv skip)	15,981,440
Text-image fusion block		channels = 2× N_c, n_heads = N_{head}, n_layers=1, d_cond = D_{cond}	9,188,480
Residual block		channels = 3× N_c, out_channels = 2× N_c (conv skip)	9,835,520
Text-image fusion block		channels = 2× N_c, n_heads = N_{head}, n_layers=1, d_cond = D_{cond}	9,188,480
up		channels = 2× N_c	3,687,040
Residual block		channels = 3× N_c, out_channels = N_c (conv skip)	3,997,120
Text-image fusion block		channels = N_c, n_heads = N_{head}, n_layers=1, d_cond = D_{cond}	2,546,240
Residual block		channels = 2× N_c, out_channels = N_c (conv skip)	2,972,480
Text-image fusion block		channels = N_c, n_heads = N_{head}, n_layers=1, d_cond = D_{cond}	2,546,240
GN		num_groups = 32, num_channels = N_c	640
SiLU		-	-
Con2D (output)		in_channels = N_c, out_channels = N_{cout}, kernel_size = 3, padding=1	2881
Residual block (channels, out_channels)			
GN		num_groups=32, num_channels=channels	2×channels
SiLU		-	0
Conv2D		in_channels=channels, out_channels=out_channels, kernel_size=3, padding=1	(3×3×channels+1)×out_channels
GN		num_groups=32, num_channels=out_channels	2×out_channels
SiLU		-	0
Dropout		p=0.0	0
Conv2D		in_channels=out_channels, out_channels=out_channels, kernel_size=3, padding=1	(3×3×out_channels+1)×out_channels
Conv2D (conv skip)		in_channels=channels, out_channels=out_channels, kernel_size=1	out_channels×(channels+1)
Text-image fusion block (channels, n_heads, n_layers, d_cond)			
GN		num_groups=32, num_channels=channels	2×channels
Conv2D		in_channels=channels, out_channels=out_channels, kernel_size=1	channels×(channels+1)
LN		normalized_shape=channels	2×channels
Self-Att	Linear (Q)	in_features=channels, out_features=channels, bias=False	channels×channels
	Linear (K)	in_features=channels, out_features=channels, bias=False	channels×channels
	Linear (V)	in_features=channels, out_features=channels, bias=False	channels×channels
	Linear	in_features=channels, out_features=channels	channels×(channels+1)
LN		normalized_shape=channels	2×channels
Cross-Att	Linear (Q)	in_features=channels, out_features=channels, bias=False	channels×channels
	Linear (K)	in_features=d_cond, out_features=channels, bias=False	d_cond×channels
	Linear (V)	in_features=d_cond, out_features=channels, bias=False	d_cond×channels
	Linear	in_features=channels, out_features=channels	channels×(channels+1)
LN		normalized_shape=channels	2×channels
FF	GeGLU	d_in=channels, d_out=4×channels	8×channels×(channels+1)
	Dropout	p=0.0	-
	Linear	in_features=4×channels, out_features=channels	channels×(4×channels+1)
Conv2D		in_channels=channels, out_channels=out_channels, kernel_size=1	channels×(channels+1)
down (channels)			
Conv2D		in_channels=channels, out_channels=channels, kernel_size=3, stride=2, padding=1	(3×3×channels+1)×channels
up (channels)			
Conv2D		in_channels=channels, out_channels=channels, kernel_size=3, padding=1	(3×3×channels+1)×channels

Supplementary Table 7 | Example input text prompts used for different tasks. DN, denoising; DCV, deconvolution, SR, super-resolution.

Task	Example text prompt
DN	Task: denoising; sample: fixed bovine pulmonary artery endothelial (BPAE) cell; structure: nuclei; fluorescence indicator: DAPI (blue); input microscope: confocal microscope with a Nikon Apo LWD 40x, 1.15 NA water immersion objective.; input pixel size: 300 x 300 nm; target microscope: confocal microscope with a Nikon Apo LWD 40x, 1.15 NA water immersion objective.; target pixel size: 300 x 300 nm.
DCV	Task: deconvolution; sample: live E. coli (Escherichia coli) cell; structure: membrane; fluorescence indicator: FM5-95 (green); input microscope: wide-field microscope with an Olympus 60x 1.42-NA Oil immersion objective. Oil refractive index 1.522.; input pixel size: 80 x 80 nm; target microscope: structured illumination microscope with an Olympus 60x 1.42NA Oil immersion objective. Oil refractive index 1.522.; target pixel size: 80 x 80 nm.
SR	Task: super-resolution with a scale factor of 2; sample: fixed COS-7 cell line; structure: clathrin-coated pits; fluorescence indicator: mEmerald (GFP); input microscope: wide-field microscope with excitation numerical aperture (NA) of 1.41, detection numerical aperture (NA) of 1.3; input pixel size: 62.6 x 62.6 nm. Nearest interpolation with a factor of 2.; target microscope: linear structured illumination microscope with excitation numerical aperture (NA) of 1.41, detection numerical aperture (NA) of 1.3; target pixel size: 31.3 x 31.3 nm.

Supplementary Table 8 | Comparison of the prediction cost between different methods. The number of parameters, the memory usage during inference, the processing time used for process image with different size (256×256, 512×512, 1024×1024) is presented. The time used for text embedding in FluoResFM is about 0.0054 s for one sample.

Model	Num. of params.	Memory usage (inference) (MiB)	Processing time (s)		
			256×256	512x512	1024x1024
CARE	922,977	542	0.002	0.018	0.048
DFCAN	1,783,041	694	0.16	0.80	2.95
UniFMIR	1,582,263	1070	0.18	1.66	4.35
FluoResFM	664,793,281	8518	0.30	2.73	7.63

References

1. Berton, G., Musgrave, K. & Masone, C. All You Need to Know About Training Image Retrieval Models. Preprint at https://doi.org/10.48550/arXiv.2503.13045 (2025).
2. Zhang, S. *et al.* A Multimodal Biomedical Foundation Model Trained from Fifteen Million Image–Text Pairs. *NEJM AI* **2**, (2025).