Supplementary Material

Benchmarking robustness of deep neural networks in semantic segmentation of fluorescence microscopy images

Table S1
Detailed training and testing configurations on Nucleus-C, Mito-C, ER-C datasets.

Dataset	Full-name	Augmentation	Image numbers (train/test)	Corrupted types	Morphology type
Nucleus-C	cell nucleus dataset	Rotation, flipping	2232 pairs (1374/858)	Noise, blurring	round spot
Mito-C	mitochondria dataset	Rotation, flipping	1200 pairs (612/588)	Noise, blurring	rod-shaped
ER-C	endoplasmic reticulum dataset	Rotation, flipping	1278 pairs (594/600)	Noise, blurring	Reticular

Table S2
Quantitative comparison of robustness before and after pixel shuffle on adversarial perturbation maps. Tested on Nucleus-C dataset.

Models	Before (%)	After (%)	Rising Rate (%)
UNet	81.64	97.08	18.91
UNet_3	68.77	78.79	14.58
Sim_UNet	66.92	80.71	20.61
FCNs	72.96	88.13	20.79
SegNet	89.97	97.58	8.46
DeepLab	83.73	98.64	17.8
PSPNet	77.37	99.07	28.05
ICNet	64.11	97.03	51.35
ViT-B_16	49.70	96.87	94.92
R50-ViT-B_16	76.49	90.42	18.21

Table S3
Quantitative comparison of robustness before and after pixel shuffle on adversarial perturbation maps. Tested on Mito-C dataset.

Models	Before (%)	After (%)	Rising Rate (%)
UNet	75.52	94.14	24.65
UNet_3	68.94	86.97	26.15
Sim_UNet	75.16	88.12	17.25
FCNs	70.80	96.80	36.72
SegNet	89.08	97.97	9.98
DeepLab	76.17	90.31	18.56
PSPNet	44.62	73.46	64.64
ICNet	53.30	89.46	67.84
ViT-B_16	53.20	94.18	77.04
R50-ViT-B_16	80.09	89.53	11.79

Table S4
Overall comparison of 10 models on three datasets in term of accuracy and five types of robustness. The models have been arranged in descending order according to 'average score'. The best result for each row is shown in bold.

Models		SegNet	FCNs	UNet	DeepLab	PSPNet	R50-ViT-B_16	ViT-B_16	ICNet	Sim_UNet	UNet_3
Conditions	Datasets										
Accuracy	Nucleus-C	86.68	**86.94**	87.50	84.55	86.16	86.42	85.61	80.37	85.75	85.53
	Mito-C	78.26	79.23	79.54	79.47	78.28	73.49	68.16	62.7	**79.90**	78.4
	ER-C	77.11	78.33	78.13	76.87	76.92	76.54	66.52	44.18	**79.88**	79.7
SNR	Nucleus-C	89.93	79.36	94.76	95.5	**96.09**	78.73	95.37	93.79	63.97	57.78
	Mito-C	82.59	**94.17**	86.18	72.01	56.71	74.68	87.65	67.21	84.63	78.66
	ER-C	**88.77**	85.6	80.48	70.27	81.34	74.24	85.75	86.55	43.65	42.42
SIB	Nucleus-C	96.51	97.07	95.5	94.07	96.59	92.63	96.61	**97.55**	91.89	90.54
	Mito-C	86.47	77.46	78.99	71.83	82.96	77.79	**89.41**	73.01	76.44	86.17
	ER-C	78.63	73.56	69.08	59.28	76.37	60.67	**83.48**	66.70	38.33	33.34
SVB	Nucleus-C	97.94	**98.64**	97.67	97.43	98.47	93.59	97.93	98.53	93.01	91.52
	Mito-C	92.41	88.96	88.74	85.82	89.65	87.62	**95.71**	87.81	86.85	92.96
	ER-C	88.24	84.90	79.33	73.77	87.05	77.41	**92.31**	85.65	50.23	46.84
FGSM	Nucleus-C	**89.97**	72.96	81.64	83.73	77.37	76.49	49.7	64.11	66.92	68.77
	Mito-C	**89.08**	70.8	75.52	76.17	44.62	80.09	53.2	53.30	75.16	68.94
	ER-C	**68.99**	54.86	68.44	55.59	47.81	68.67	47.84	45.58	31.69	38.47
IFGSM	Nucleus-C	63.15	66.3	53.90	**76.09**	63.86	50.95	27.3	50.92	48.74	50.13
	Mito-C	**82.04**	63.7	48.09	56.98	58.87	48.9	38.77	40.43	49.42	41.36
	ER-C	46.58	**50.74**	33.04	47.84	35.38	47.91	39.27	35.38	28.11	30.14
Averaged Score		**82.41**	77.98	76.47	75.4	74.14	73.71	72.26	68.54	65.25	64.54
Corrupt-biased Score		**84.23**	80.93	79.75	77.27	78.04	76.02	78.17	72.59	68.3	67.52
Adv-biased Score		**80.19**	75.09	73.42	73.87	70.51	71.71	65.81	63.36	63.71	63.08

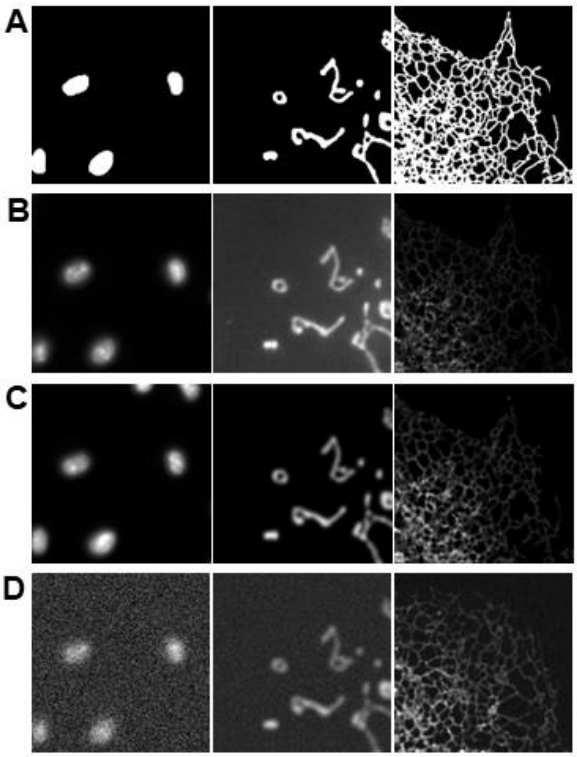

Fig. S1. Samples of the image synthesis workflow. (A) Binary masks. (B) Initial synthetic images. (C) Denoised synthetic images. (D) Output images with controlled degradations.

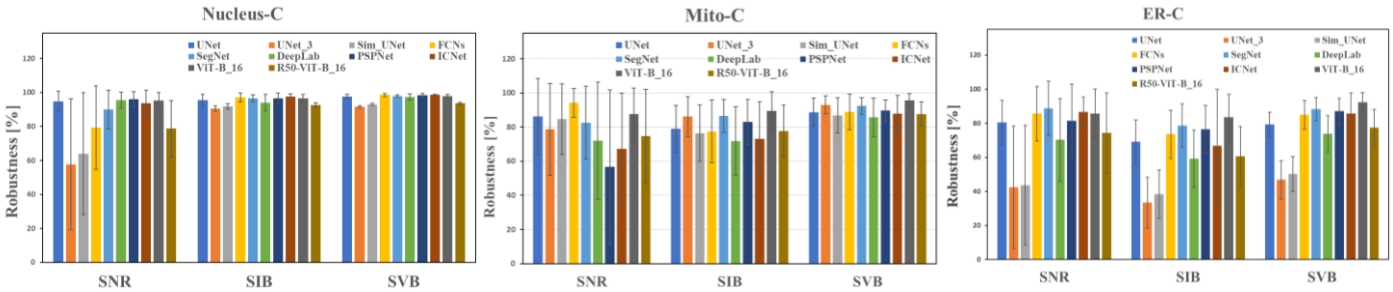

Fig. S2. Corruption robustness of three types on three datasets.
Error bars indicate standard deviation. The sample size for each experiment is 500.

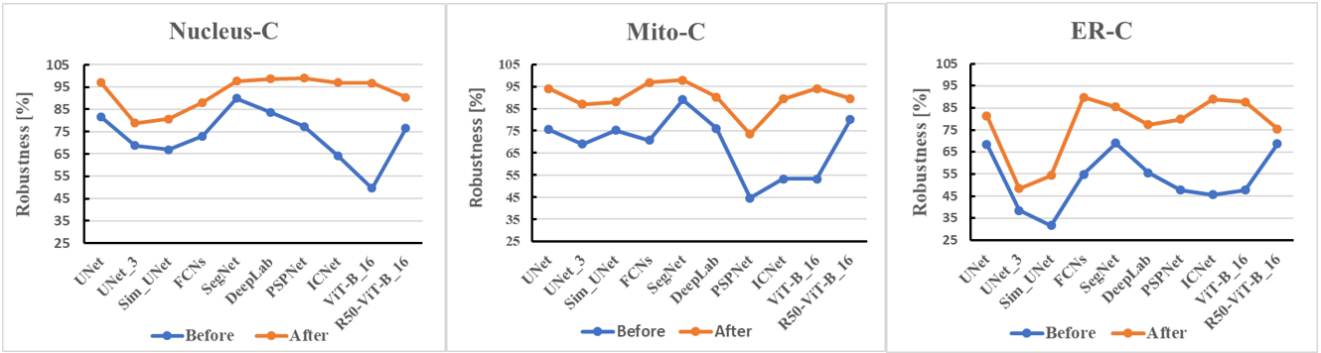

Fig. S3. Comparison of robustness before and after randomized shuffling of pixels of adversarial residual maps

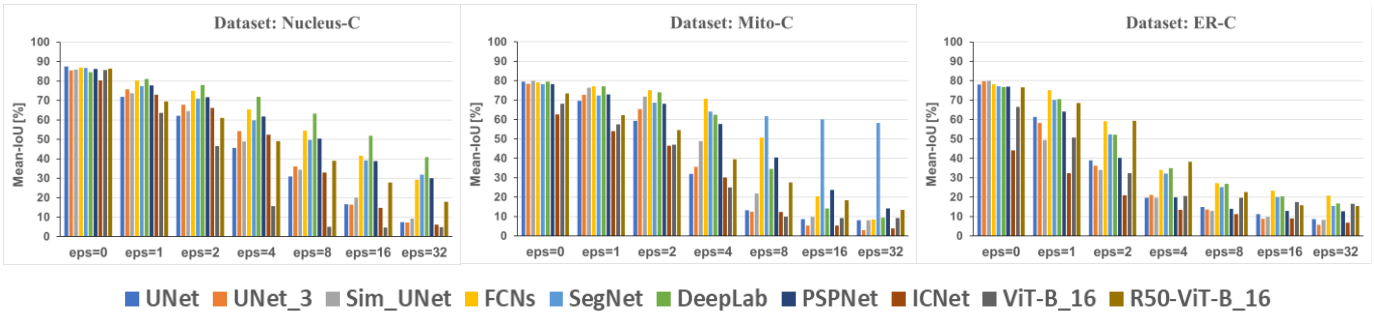

Fig. S4. Model accuracy under different levels of IFGSM attacks. eps: ε controls perturbation.

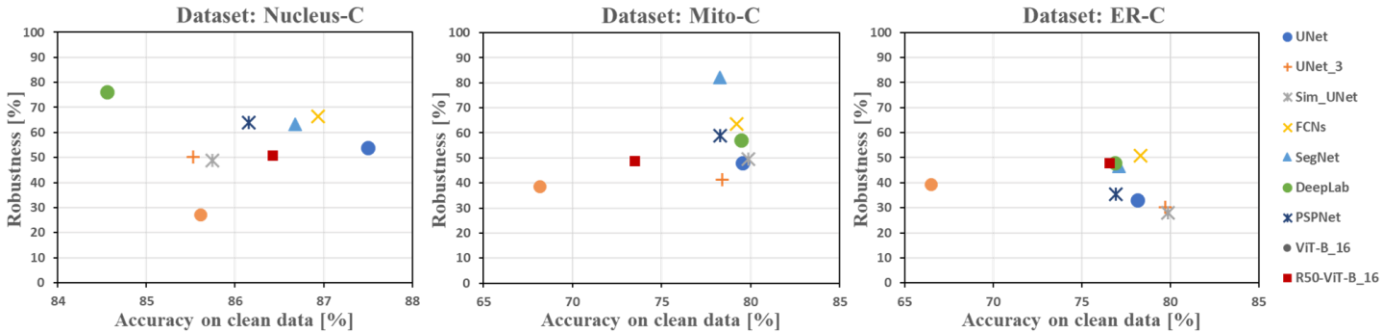

Fig. S5. Adversarial robustness of models against IFGSM attacks on three datasets . Results for ICNet are not shown because its very low accuracy skews the results of other models.

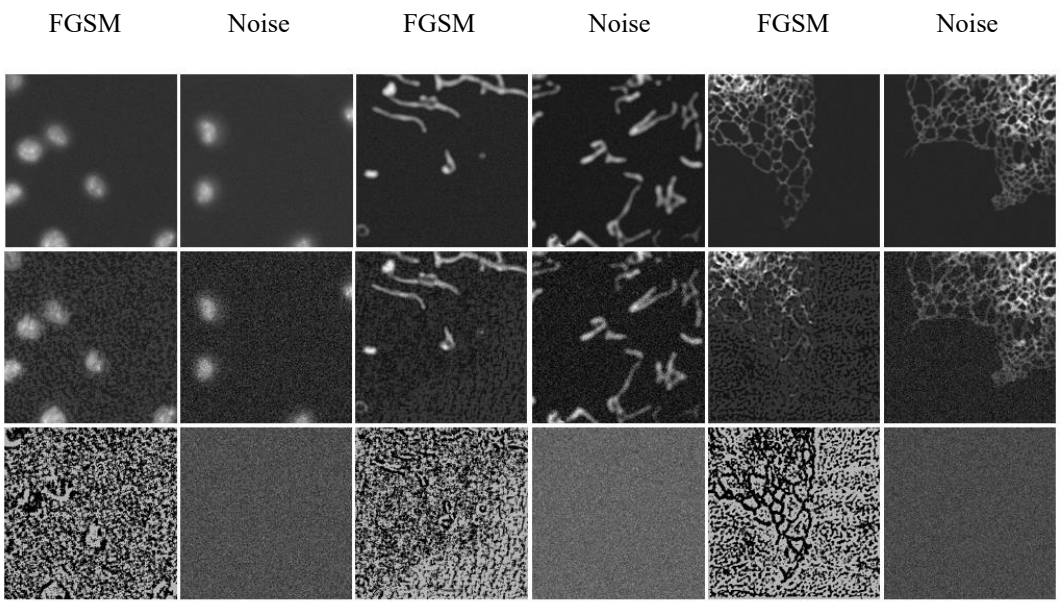

Fig. S6. Comparison of adversarial attack images and noise corrupted images. From the first row to the bottom: original images, degraded images and corresponding residual maps.

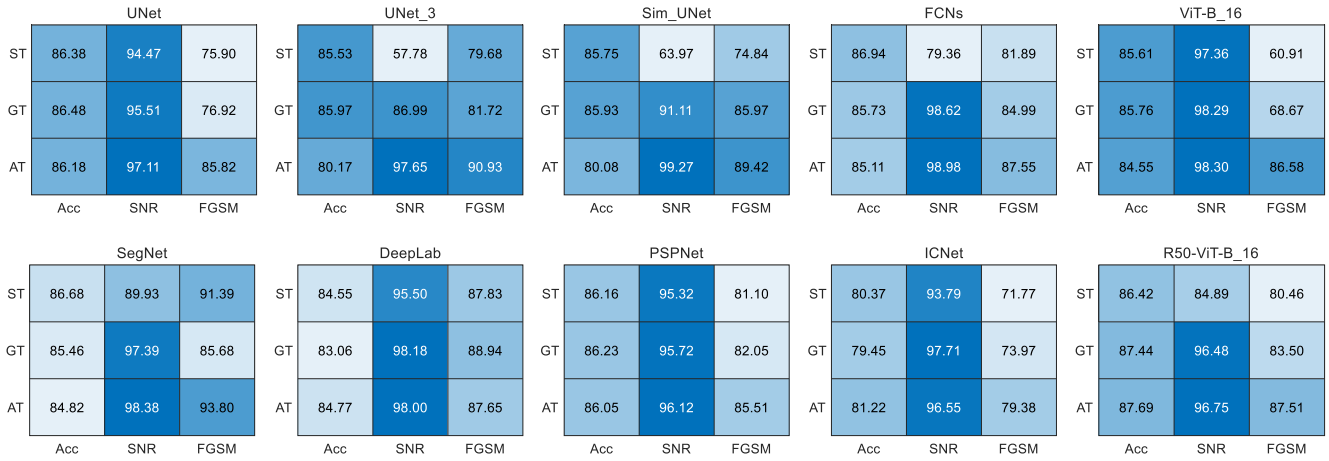

Fig. S7. Robustness with three types of training on Nucleus-C dataset. ST: standard training, GT: Gaussian noise augmentation training, AT: adversarial training, Acc: mIoU on clean data.

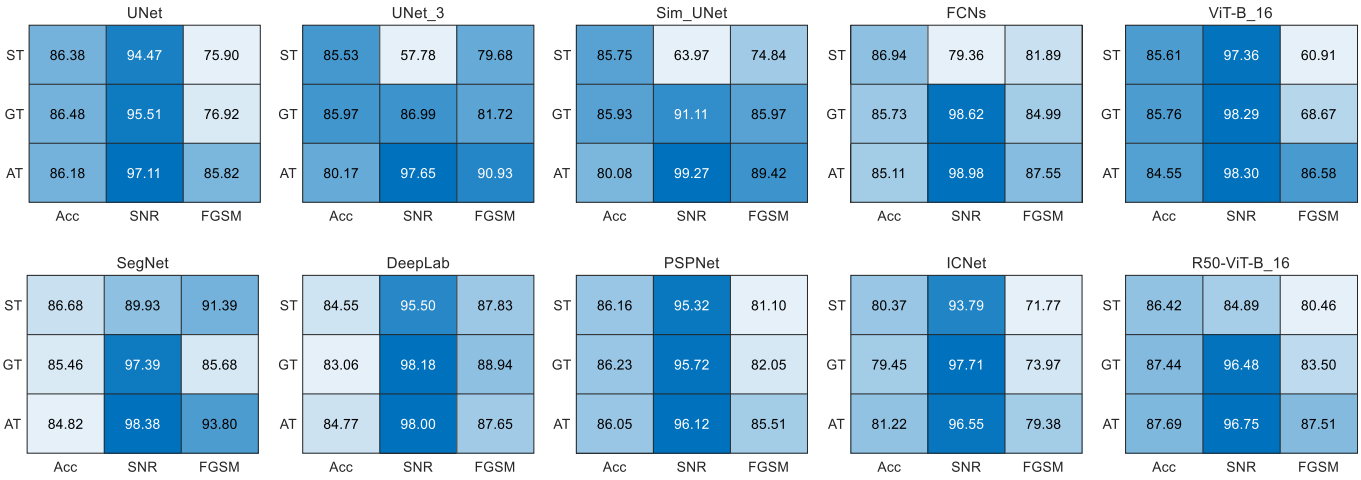

Fig. S8. Robustness with three types of training on ER-C dataset. ST: standard training, GT: Gaussian noise augmentation training, AT: adversarial training, Acc: mIoU on clean data.

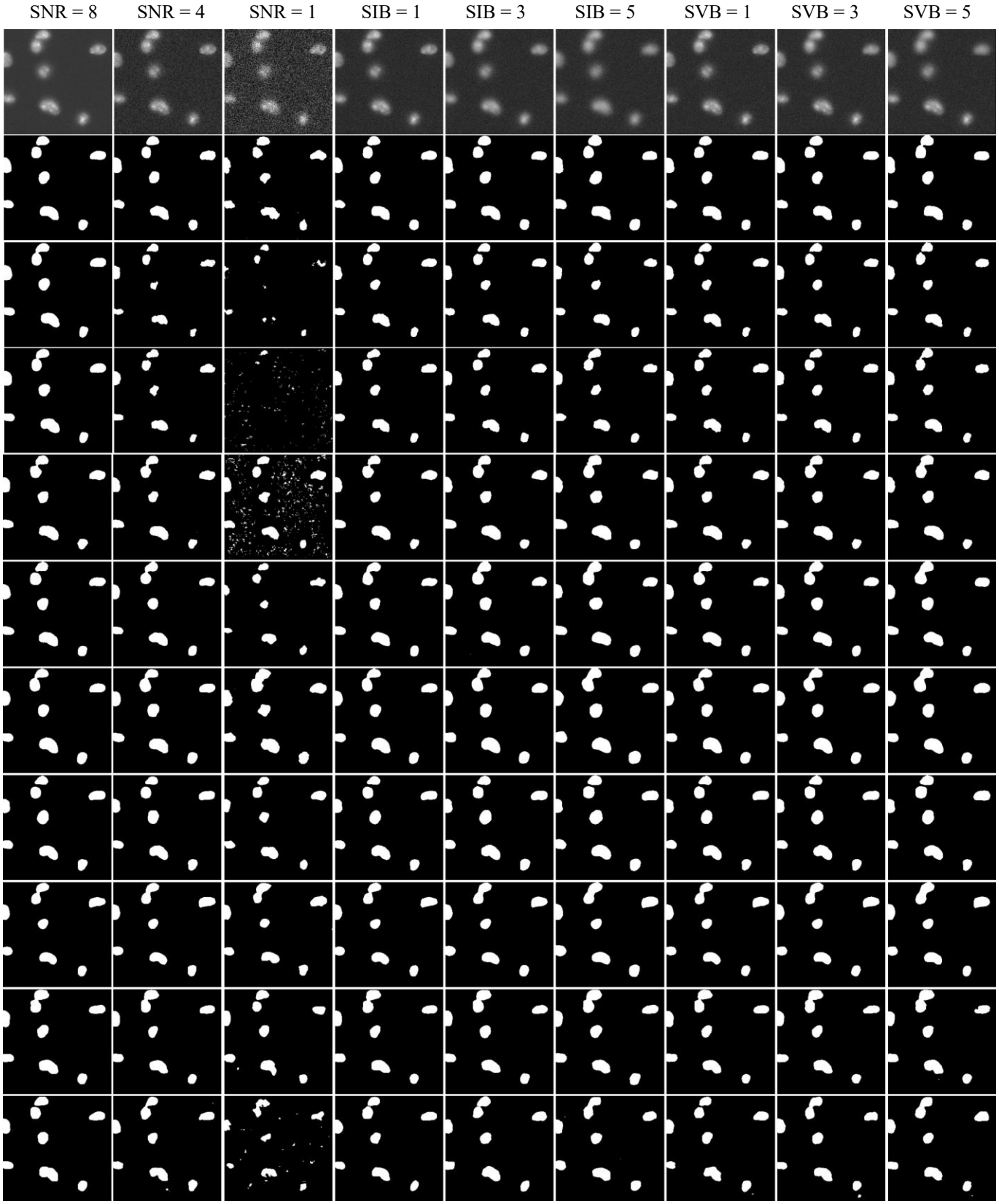

Fig. S9. Three types of corruption SNR, SIB, SVB on Nucleus-C dataset and corresponding segmentation results. Corrupted images to be tested are list in the first row, images with SNR = 8 are clean data. From the second row to the last, they are segmentation results of UNet, UNet_3, Sim_UNet, FCNs, SegNet, DeepLab, PSPNet, ICNet, ViT-B_16, R50-ViT-B_16.

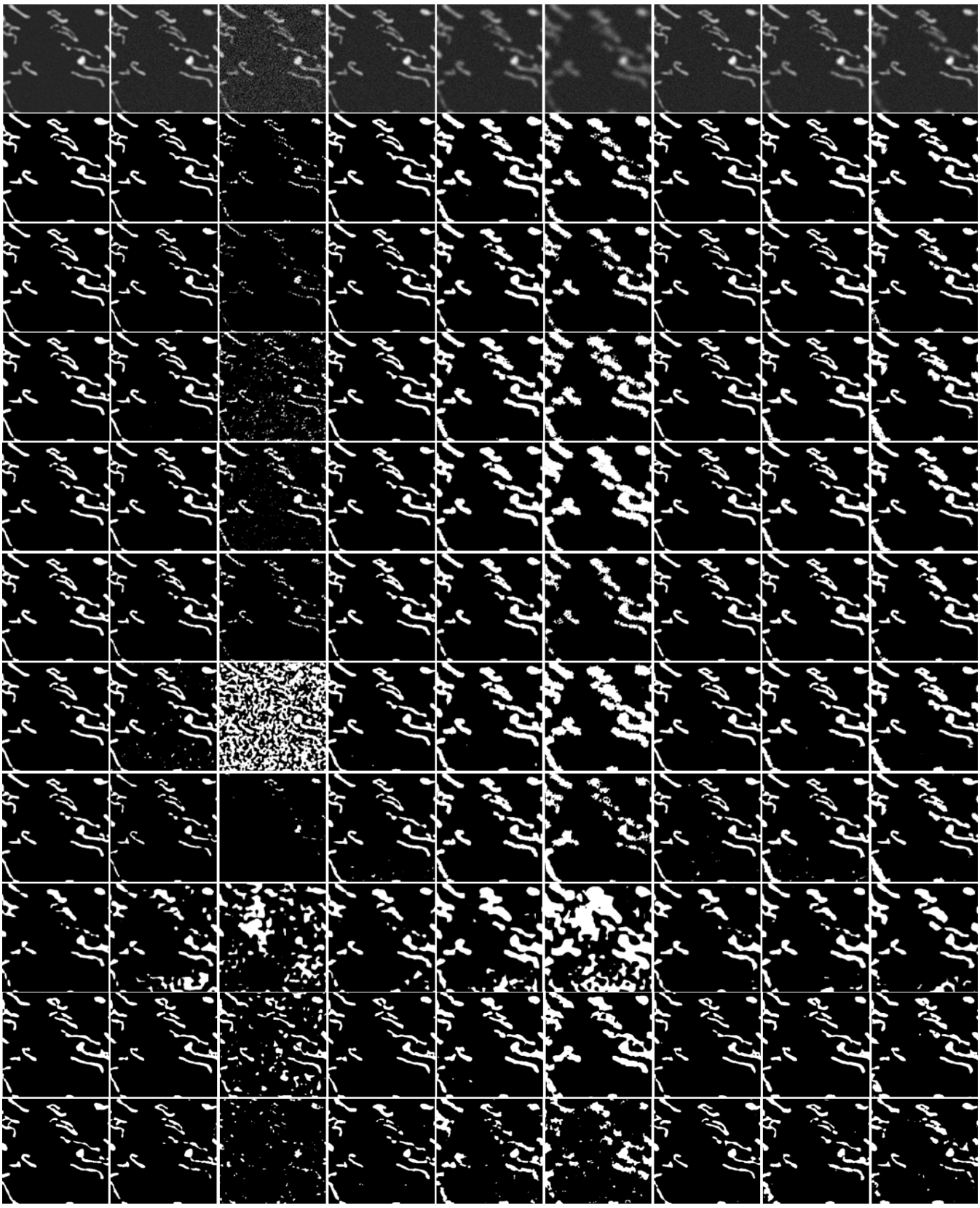

Fig. S10. Three types of corruption SNR, SIB, SVB on Mito-C dataset and corresponding segmentation results. Corrupted images to be tested are list in the first row, images with SNR = 8 are clean data. From the second row to the last, they are segmentation results of UNet, UNet_3, Sim_UNet, FCNs, SegNet, DeepLab, PSPNet, ICNet, ViT-B_16, R50-ViT-B_16.

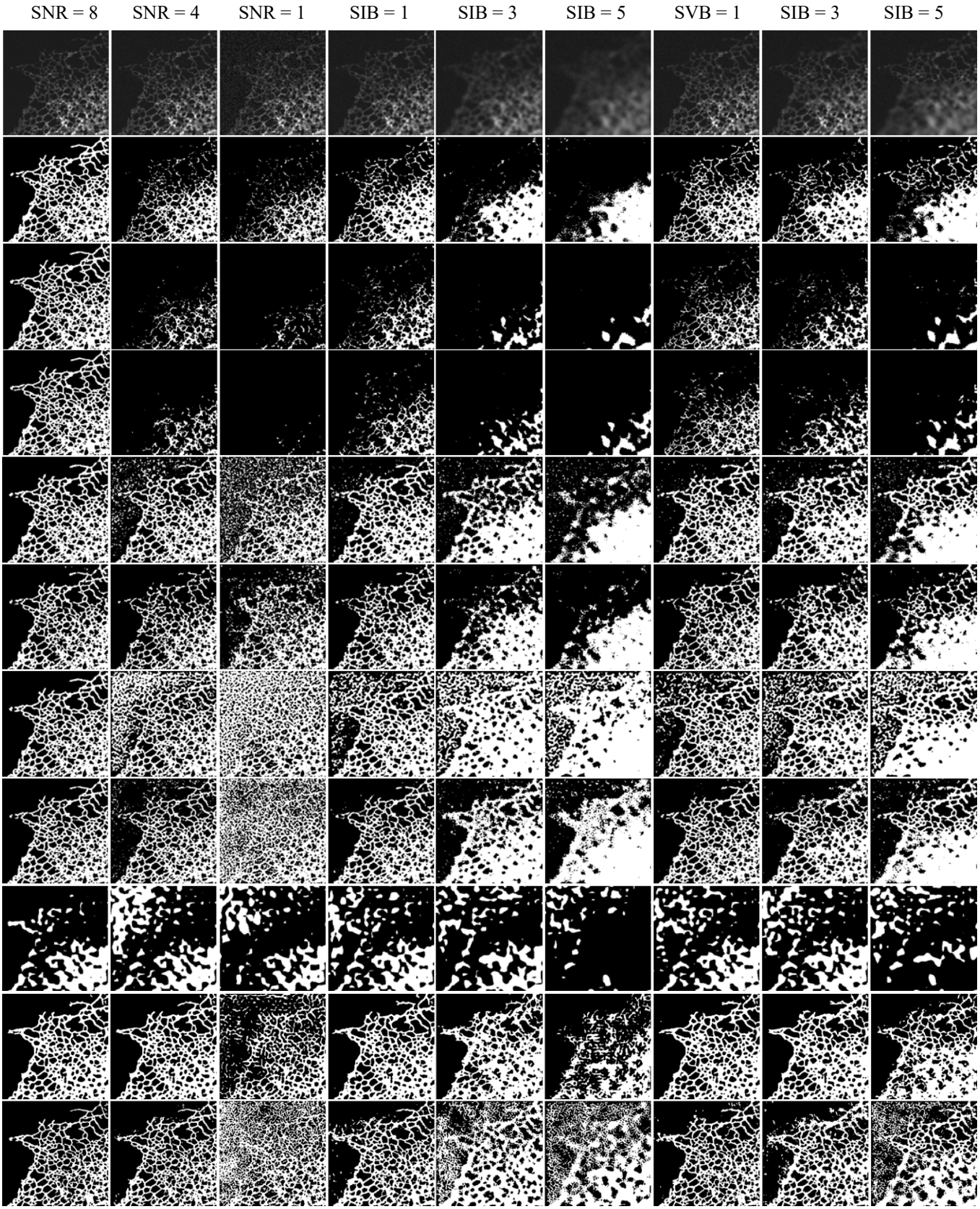

Fig. S11. Three types of corruption SNR, SIB, SVB on ER-C dataset and corresponding segmentations. Corrupted images to be tested are list in the first row, images with SNR = 8 are clean data. From the second row to the last, they are segmentations of UNet, UNet_3, Sim_UNet, FCNs, SegNet, DeepLab, PSPNet, ICNet, ViT-B_16, R50-ViT-B_16.

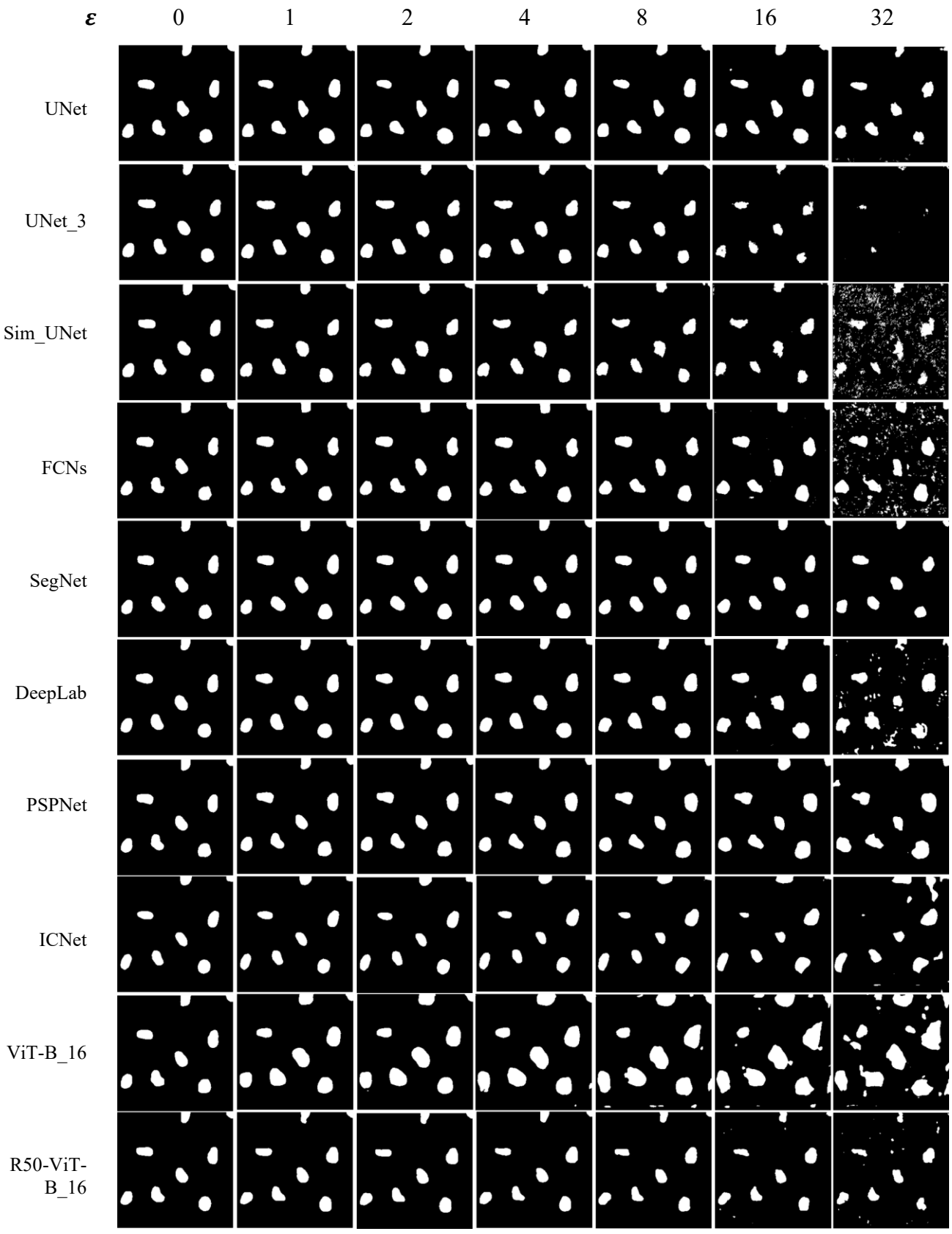

Fig. S12. Segmentation results on Nucleus-C dataset perturbated by FGSM attack. ε controls the level of attack.

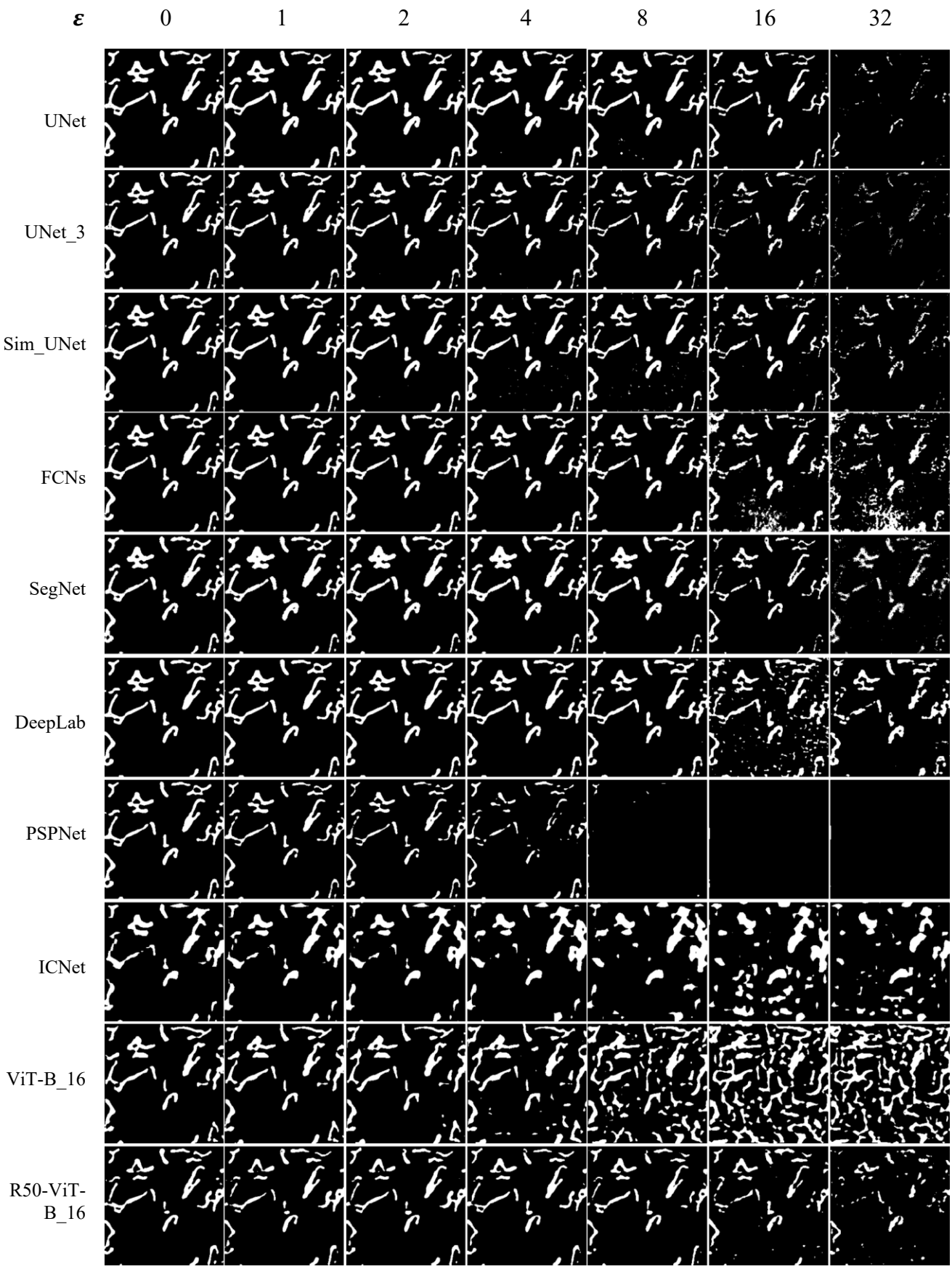

Fig. S13. Segmentation results on Mito-C dataset perturbated by FGSM attack. ε controls the level of attack.

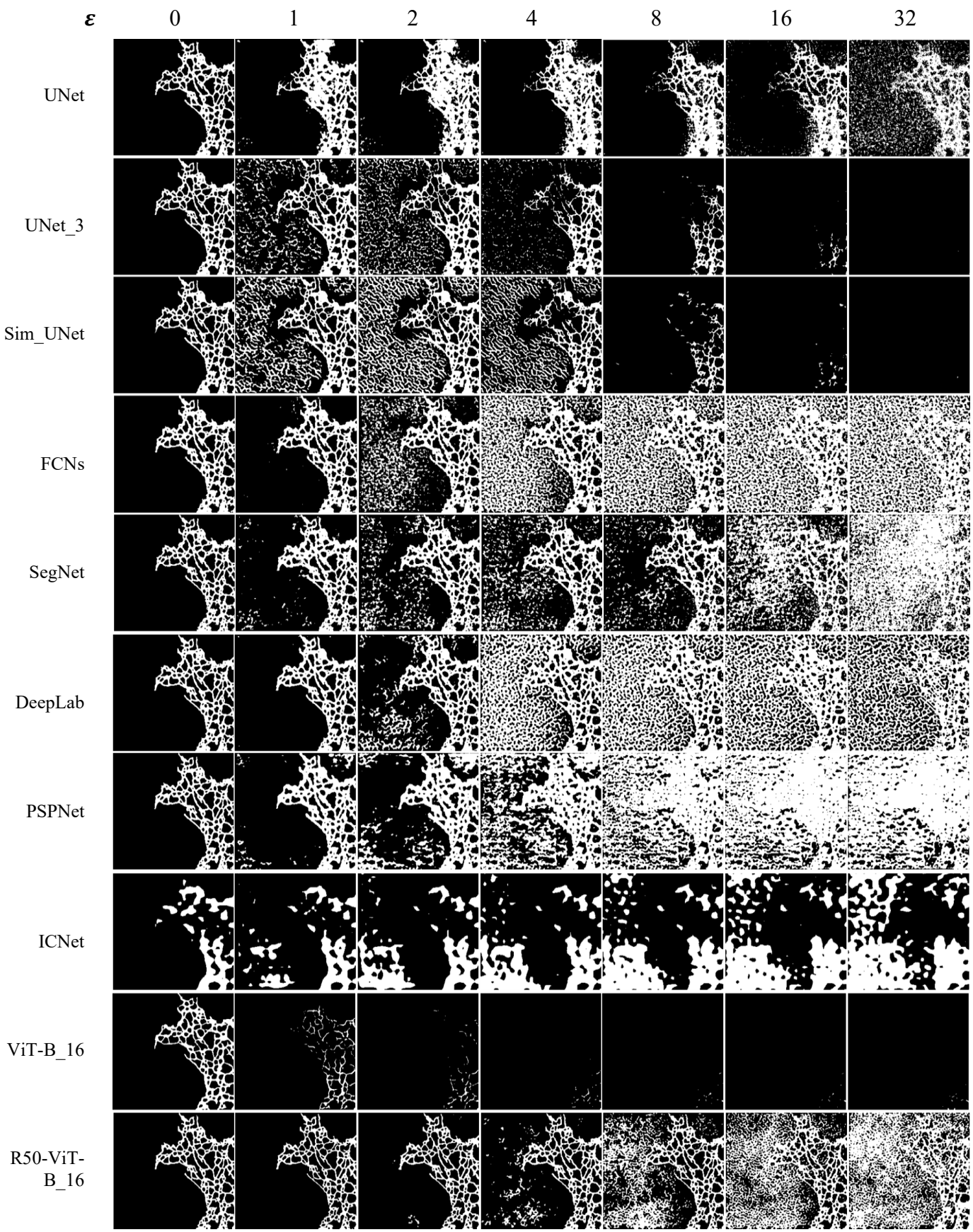

Fig. S14. Segmentation results on ER-C dataset perturbated by IFGSM attack. ε controls the level of attack.

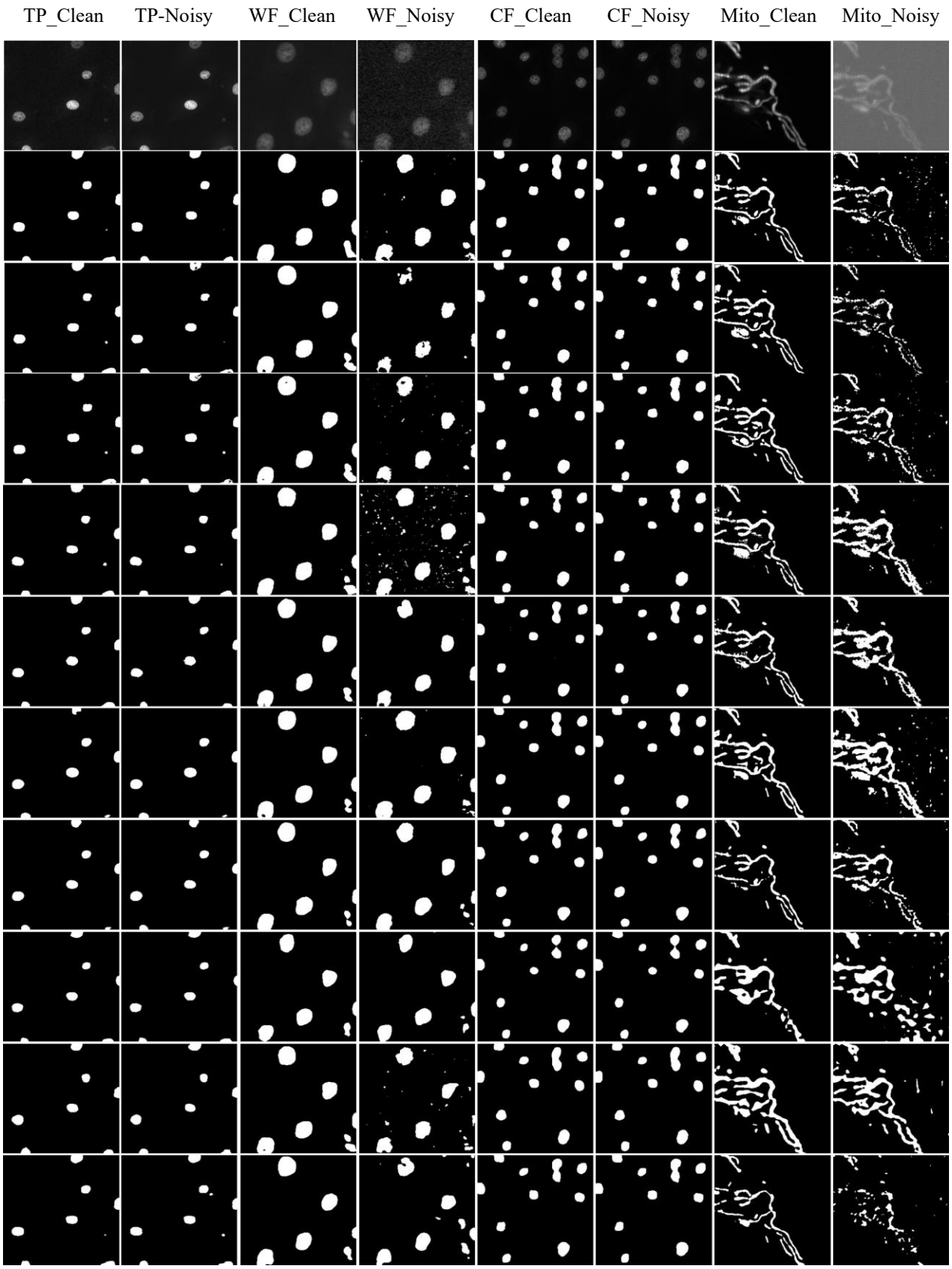

Fig. S15. Segmentation results of real FM images. CF: confocal (nucleus), TP: twophoton (nucleus), WF: widefield (nucleus). From the second row to the last: results of UNet, UNet_3, Sim_UNet, FCNs, SegNet, DeepLab, PSPNet, ICNet, ViT-B_16, R50-ViT-B_16.