Toward Generalizable Whole Brain Representations with High-Resolution Light-Sheet Data

Supplementary Material

6. Supplementary

6.1. Data Acquisition

6.1.1. Sample Processing

To generate raw datasets for each marker, we used *Swiss Webster* mice (*8 wks+*, *Females*) for immunolabeling experiments. The parvalbumin dataset was generated using a *B6 Pvalb-IRES-Cre::tdTomato* mouse (*8 wks*, *Female*). The following primary antibodies were used for immunolabeling: anti-NeuN CST #36662 (*6 μg*), anti-TH Biolegend #818001 (*6 μg*), anti-c-Fos Abcam #ab214672 (*3.5 μg, no special manipulation*), anti-IBA1 CST #79394 (*6 μg*), and anti-GFAP Invitrogen #13-0300 (*6 μg*).

For tissue processing, all incubations were performed in 20 mL of the respective solution with light shaking unless otherwise specified. Mice were first transcardically perfused with 4% Paraformaldehyde (PFA), then extracted brains were fixed in 4% PFA for additional 24 hours at 4°C. For SHIELD preservation[34], samples were incubated in SHIELD-OFF solution for 3 days at 4 °C, then in SHIELD-ON solution for 1 day at 37 °C. Samples were then delipidated in Clear+ Delipidation Buffer for 7 d at 45 °C. Whole brain immunolabeling steps were performed using eFLASH technology[44], which integrates stochastic electrotransport[23], implemented on a SmartBatch+ device (LifeCanvas Technologies). Indirect immunolabeling was performed using 1:2 molar ratio between primary and secondary antibodies. Primary immunolabeling was performed using the RADIANT Buffer System (LifeCanvas Technologies), followed by secondary labeling using the SmartBatch+ Secondary Buffer System (LifeCanvas Technologies), with the device presets Labeling 1 at 24 hours and Labeling 2 at 12 hours, respectively.

6.1.2. Imaging

For optical clearing, samples were incubated in 50% EasyIndex in denionized water, then in 100% EasyIndex (RI = 1.52; LifeCanvas Technologies) both for 1 day at 37 °C. Refractive index matched samples were imaged using a SmartSPIM axially swept light-sheet microscope (LifeCanvas Technologies) at 3.6× (0.2 NA) magnification. Regarding potential destriping artifacts, we have carefully tuned the FFT–wavelet filter over hundreds of samples to prioritize signal preservation, even if minor residual striping remains, and observe improved image quality[39].

6.2. Baseline Models

6.2.1. Cell location prediction

The valid detection satisfies Equation (2) with the following algorithm.

Algorithm 2: Model with Location Prediction

Input : Detection Model M, threshold $\tau \in [0, 1]$
Output : Location prediction model M_{loc}
Parameters: $d_{min} = 3$ (minimum distance between peaks)

$I_{img} \leftarrow$ Volumetric Image Input from Model M
$I_{idx} \leftarrow$ Batch index
$H \leftarrow$ Model output; 3D probability heatmap
$k \leftarrow 2 \times d_{min} + 1$
$H_{max} \leftarrow \texttt{MaxPool3D}(H, \text{kernel} = (k, k, k), \text{stride} = (1, 1, 1), \text{padding} = \text{"valid"})$
$H_{max} \leftarrow \texttt{ZeroPad3D}(H_{max}, \text{padding} = d_{min})$
$L \leftarrow \texttt{FindMaxima}(H, H_{max}, I_{idx}, \tau)$
$M_{loc} \leftarrow \texttt{Model}(I_{img}, I_{idx}, L)$
return M_{loc}

6.2.2. Cross-dataset performance of baseline models

Table 4 shows comprehensive evaluation of each baseline model across all six datasets. In general, models performed best on their corresponding cell type, confirming dataset-specific specialization. However, notable exceptions exist: the GFAP model performed poorly on its own dataset (F1=0.33), where the Iba1 model substantially outperformed it (F1=0.61), likely due to GFAP having the fewest annotations (2,530 total GT cells) while the Iba1 model benefits from a significantly larger training set (6,170 GT cells). Similarly, for some PV and cFos regions, non-matched models occasionally achieved comparable or higher F1 scores. Bold indicates evaluation on the matched dataset; underline indicates the highest-performing model for each region.

6.2.3. Preliminary evaluations of segmentation performance

The downstream analyses enabled by the CANVAS dataset include regional cell density estimation, multi-channel co-expression analysis, and spatial proximity analysis between cell types. For these tasks, cell centroid annotations are generally sufficient and are substantially faster and more reliable to generate (particularly in large 3D datasets) than full instance segmentations. This is especially true for densely packed, irregularly shaped cells such as GFAP$^+$ astrocytes, where overlapping processes make instance boundaries ambiguous and segmentation error-prone (Fig. 4d).

Table 4. Performance metrics (Accuracy and F1 Score) for six cell-type models across regions. (**Bold** indicates evaluation on matched data; underline indicates the highest-performing model.)

Cell Type	Region	cFos Model		NeuN Model		TH Model		PV Model		GFAP Model		Iba1 Model	
		Accuracy	F1_Score	Accuracy	F1_Score	Accuracy	F1_Score	Accuracy	F1_Score	Accuracy	F1_Score	Accuracy	F1_Score
cFos	**Total**	**0.64**	**0.78**	0.62	0.76	0.15	0.26	0.59	0.74	0.17	0.29	0.58	0.74
	Train Set	**0.63**	**0.77**	0.63	0.78	0.14	0.25	0.60	0.75	0.16	0.28	0.58	0.73
	region_1	**0.66**	**0.80**	0.62	0.76	0.13	0.23	0.60	0.75	0.18	0.31	0.56	0.72
	region_2	**0.72**	**0.84**	0.67	0.80	0.15	0.25	0.67	0.80	0.11	0.19	0.58	0.73
	region_3	**0.53**	**0.69**	0.62	0.77	0.16	0.27	0.53	0.70	0.19	0.32	0.60	0.75
	Test Set	**0.64**	**0.78**	0.60	0.75	0.16	0.27	0.59	0.74	0.17	0.30	0.59	0.74
	region_4	**0.66**	**0.80**	0.60	0.75	0.15	0.25	0.55	0.71	0.18	0.30	0.57	0.72
	region_5	**0.74**	**0.85**	0.62	0.77	0.15	0.26	0.73	0.84	0.12	0.22	0.62	0.76
	region_6	**0.54**	**0.70**	0.58	0.73	0.18	0.31	0.52	0.69	0.21	0.35	0.59	0.74
NeuN	**Total**	0.01	0.02	**0.67**	**0.81**	0.12	0.21	0.57	0.73	0.02	0.04	0.40	0.57
	Train Set	0.01	0.02	**0.65**	**0.79**	0.11	0.20	0.57	0.72	0.02	0.04	0.38	0.55
	region_1	0.00	0.00	**0.63**	**0.77**	0.13	0.23	0.55	0.71	0.02	0.03	0.41	0.58
	region_2	0.07	0.13	**0.53**	**0.69**	0.13	0.23	0.38	0.55	0.06	0.11	0.31	0.48
	region_3	0.01	0.02	**0.70**	**0.83**	0.09	0.17	0.63	0.77	0.02	0.04	0.37	0.54
	Test Set	0.01	0.02	**0.69**	**0.82**	0.12	0.22	0.58	0.73	0.02	0.04	0.41	0.59
	region_4	0.00	0.00	**0.73**	**0.84**	0.13	0.23	0.56	0.72	0.02	0.03	0.38	0.55
	region_5	0.09	0.16	**0.51**	**0.67**	0.13	0.23	0.40	0.57	0.06	0.11	0.32	0.48
	region_6	0.01	0.02	**0.70**	**0.83**	0.12	0.21	0.64	0.78	0.02	0.04	0.48	0.64
TH	**Total**	0.02	0.04	0.25	0.41	**0.40**	**0.57**	0.11	0.20	0.08	0.14	0.16	0.28
	Train Set	0.01	0.02	0.25	0.40	**0.42**	**0.59**	0.06	0.12	0.07	0.13	0.16	0.28
	region_1	0.01	0.01	0.22	0.36	**0.40**	**0.57**	0.02	0.03	0.09	0.16	0.13	0.23
	region_2	0.00	0.01	0.11	0.21	**0.35**	**0.51**	0.03	0.06	0.02	0.04	0.07	0.13
	region_3	0.04	0.07	0.54	0.70	**0.66**	**0.79**	0.21	0.34	0.11	0.19	0.29	0.45
	Test Set	0.03	0.06	0.26	0.41	**0.38**	**0.55**	0.16	0.28	0.09	0.16	0.16	0.27
	region_4	0.01	0.02	0.18	0.31	**0.35**	**0.51**	0.03	0.07	0.08	0.14	0.09	0.17
	region_5	0.01	0.03	0.08	0.14	**0.29**	**0.45**	0.04	0.07	0.01	0.03	0.03	0.05
	region_6	0.08	0.14	0.64	0.78	**0.71**	**0.83**	0.55	0.71	0.20	0.34	0.36	0.53
PV	**Total**	0.16	0.28	0.81	0.89	0.52	0.68	**0.46**	**0.63**	0.12	0.21	0.45	0.62
	Train Set	0.17	0.29	0.81	0.89	0.50	0.67	**0.42**	**0.59**	0.11	0.20	0.48	0.65
	region_1	0.18	0.30	0.67	0.81	0.66	0.79	**0.50**	**0.66**	0.12	0.22	0.33	0.50
	region_2	0.00	0.00	0.78	0.87	0.33	0.49	**0.23**	**0.37**	0.03	0.07	0.43	0.60
	region_3	0.40	0.58	0.99	1.00	0.96	0.98	**0.96**	**0.98**	0.31	0.47	0.86	0.93
	Test Set	0.16	0.28	0.81	0.89	0.53	0.70	**0.51**	**0.67**	0.13	0.23	0.42	0.60
	region_4	0.15	0.27	0.65	0.79	0.58	0.73	**0.53**	**0.70**	0.12	0.22	0.27	0.42
	region_5	0.00	0.01	0.79	0.88	0.39	0.56	**0.35**	**0.52**	0.04	0.08	0.37	0.54
	region_6	0.39	0.56	0.99	1.00	0.95	0.98	**0.97**	**0.98**	0.36	0.53	0.87	0.93
GFAP	**Total**	0.00	0.00	0.03	0.05	0.02	0.04	0.00	0.01	**0.20**	**0.33**	0.44	0.61
	Train Set	0.00	0.00	0.01	0.03	0.01	0.03	0.00	0.01	**0.19**	**0.32**	0.41	0.58
	region_1	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	**0.20**	**0.34**	0.48	0.65
	region_2	0.00	0.00	0.03	0.05	0.03	0.06	0.00	0.00	**0.15**	**0.26**	0.31	0.48
	region_3	0.00	0.00	0.13	0.23	0.12	0.21	0.00	0.00	**0.19**	**0.31**	0.23	0.37
	Test Set	0.00	0.00	0.05	0.09	0.03	0.06	0.00	0.00	**0.22**	**0.35**	0.49	0.66
	region_4	0.00	0.00	0.00	0.01	0.01	0.03	0.00	0.01	**0.18**	**0.30**	0.48	0.65
	region_5	0.00	0.00	0.12	0.21	0.05	0.10	0.00	0.01	**0.23**	**0.37**	0.51	0.67
	region_6	0.00	0.01	0.08	0.15	0.05	0.10	0.00	0.00	**0.28**	**0.44**	0.50	0.67
Iba1	**Total**	0.01	0.03	0.27	0.43	0.39	0.56	0.03	0.06	0.28	0.43	**0.69**	**0.81**
	Train Set	0.01	0.01	0.24	0.39	0.35	0.52	0.02	0.05	0.27	0.42	**0.64**	**0.78**
	region_1	0.02	0.03	0.41	0.59	0.47	0.64	0.05	0.09	0.30	0.46	**0.72**	**0.84**
	region_2	0.00	0.01	0.18	0.30	0.37	0.54	0.02	0.03	0.21	0.35	**0.56**	**0.72**
	region_3	0.00	0.00	0.12	0.22	0.21	0.34	0.01	0.02	0.29	0.45	**0.65**	**0.79**
	Test Set	0.02	0.04	0.30	0.46	0.43	0.60	0.03	0.07	0.29	0.44	**0.73**	**0.85**
	region_4	0.01	0.03	0.43	0.60	0.46	0.63	0.06	0.11	0.32	0.48	**0.74**	**0.85**
	region_5	0.01	0.03	0.32	0.48	0.46	0.63	0.03	0.05	0.23	0.38	**0.65**	**0.78**
	region_6	0.03	0.06	0.18	0.31	0.37	0.54	0.02	0.03	0.30	0.46	**0.83**	**0.91**

However, some biomarkers cannot be adequately represented by centroids alone. For example, β-amyloid plaques lack a meaningful "center" and require segmentation masks to capture their spatial extent and morphology. Segmentation also enables more detailed morphological characterization of cellular structures. To support this type of analysis, we are developing a new β-amyloid segmentation dataset based on ThioflavinS-stained adult AD mouse brains (*5xFAD*) that will serve as a segmentation benchmark in future releases of CANVAS.

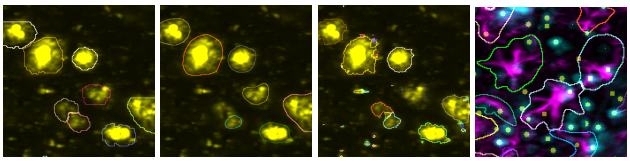

Figure 4. Example β-amyloid (yellow) segmentation results for (a) ground truth, (b) μSAM, and (c) CellPose-SAM. (d) μSAM segmentation overlaid with GFAP (magenta), nuclear stain (cyan), and detected nuclei cell centers (yellow).

As a preliminary evaluation, we compared two recent SAM-based segmentation models (Cellpose-SAM [32] and μSAM [1]) on β-amyloid plaques within a subsection of this dataset. Performance was evaluated using panoptic quality (PQ) [24], which jointly measures recognition quality (object-level F1) and segmentation quality (mean IoU over matched instances). A $200 \times 1000 \times 1000$ (z, y, x) region was cropped for evaluation, and image intensities were clipped and normalized to the 99.5th percentile.

Qualitatively, μSAM was able to capture large β-amyloid plaques, whereas Cellpose-SAM frequently over-segmented plaques into smaller fragments (Fig. 4). This resulted in lower PQ for Cellpose-SAM (0.05) compared to μSAM (0.19). Although these models were not fine-tuned on our domain-specific data, the results suggest that SAM-based segmentation models already show promising performance on LSFM datasets. In future work, we plan to fine-tune models such as μSAM on the CANVAS dataset to further improve segmentation performance and establish stronger benchmarks for 3D biomarker segmentation.

6.3. Representation learning with CANVAS

6.3.1. Model architecture

For training the autoencoder part of the 3D-MAE model, we used a standard Vision Transformer architecture adapted for 3D volumetric data with the configuration shown in Table 5, and hyperparameter settings shown in Table 6.

6.3.2. 3D-MAE model training results

Table 7 shows the 3-D MAE training results. Overall, lower mask ratios (0.15–0.55) outperformed the commonly used ratio of 0.75 reported in the original paper. Additionally, both Structural Similarity Index Measure (SSIM) and Peak Signal-to-Noise Ratio (PSNR) values for the best models are lower than those typically observed when training on natural images such as ImageNet. This is expected, as 3-D brain microscopy data contain sparse biological structures and SSIM is highly sensitive to structural patterns; with most patches being empty or nearly empty, SSIM provides limited meaningful signal. Nevertheless, the goal of this training is to learn useful representations for downstream tasks rather than to maximize reconstruction metrics.

Among the marker-specific trainings, NeuN achieved the lowest loss, consistent with its compact and relatively uniform cellular morphology. In contrast, GFAP reported the highest loss, reflecting the complexity and elongated processes of astrocytes, which make reconstruction more challenging. The all-markers model performed within 15% of the best single-marker models despite capturing roughly six times greater morphological diversity.

Table 5. 3D-MAE Model Configuration

Component	Configuration	Parameters
Encoder		10.75M
Hidden dimension	384	
Number of layers	6	
Attention heads	6	
MLP ratio	4	
Decoder		1.90M
Hidden dimension	192	
Number of layers	4	
Attention heads	3	
Total		**12.65M**

Table 6. Complete Hyperparameter Configuration

Parameter	**Value**
Architecture	
Crop sizes	16×32×32, 24×48×48, 32×64×64
Patch sizes	4×8×8, 6×12×12, 8×16×16
Mask ratios	0.15, 0.35, 0.55, 0.75
Optimization	
Optimizer	AdamW
Learning rate (η)	$1.5e^{-4}$
Weight decay	0.05
Batch size	64
Epochs	700
LR schedule	Cosine annealing
Min LR	$0.01 \times \eta$
Gradient clipping	$\|g\|_2 \leq 1.0$
Data Augmentation	
Intensity rescaling	[0, 1] normalization
Random flip	0.5
Random rotation	$\pm 10°$
Gaussian noise	$\sigma = 0.01$
Content-Aware Weighting	
Background weight	$w_{\text{bg}} = 1.0$
Cell weight	$w_{\text{cell}} \approx 10.0$

6.3.3. Content-Aware Weighting Ablation

To validate our content-aware reconstruction weighting ($w_{\text{cell}} = 10 \times w_{\text{bg}}$), we compared it with uniform weighting. Under uniform weighting, the model over-optimizes for trivial background reconstruction due to the high sparsity of our volumetric microscopy data. In contrast, content-aware weighting based on patch variance concentrates learning on cellular regions while preserving spatial context, yielding

Table 7. 3-MAE training results. Best configuration (16×32×32 / 4×8×8).

Marker	Config	Mask	Loss ↓	PSNR ↑	SSIM ↑
NeuN	16×32×32 / 4×8×8	0.15	0.0061	12.00	0.256
cFos	16×32×32 / 4×8×8	0.55	0.0082	14.55	0.291
TH	16×32×32 / 4×8×8	0.55	0.0143	13.38	0.295
PV	16×32×32 / 4×8×8	0.15	0.0087	12.69	0.199
IBA1	16×32×32 / 4×8×8	0.15	0.0103	15.61	0.194
GFAP	16×32×32 / 4×8×8	0.55	0.0194	13.74	0.360
all_markers	16×32×32 / 4×8×8	0.15	0.0070	14.68	0.172

5.1% and 10.8% improvements in PSNR and SSIM, respectively, for the NeuN data set (Table 8).

Table 8. Content-Aware Weighting Ablation (NeuN, 16×32×32 / 4×8×8, m=0.15)

Weighting Strategy	Final Loss	PSNR (dB)	SSIM
Uniform ($w_i = 1.0$ for all)	0.0068	11.42	0.231
Content-aware ($w_i \in [1, 10]$)	**0.0061**	**12.00**	**0.256**
Relative improvement	+11.5%	+5.1%	+10.8%

6.3.4. Qualitative Results and Visualizations

Representative reconstruction examples from the 3D-MAE training (Figure 5, Figure 6, Figure 7, Figure 8, Figure 9, Figure 10, Figure 11, Figure 12) illustrate the model's ability to reconstruct cellular morphology from masked patches. Each figure shows all Z-slices concatenated horizontally, as indicated by the arrow, and arranged in four rows: the original volume, the masked input with yellow dotted borders marking masked areas, the reconstructed output, and the merged view combining visible and reconstructed regions.

Across markers, NeuN exhibits the sharpest reconstructions owing to its compact and uniform morphology, while IBA1 preserves fine ramified processes despite their complexity. GFAP shows some blurring along elongated processes but retains overall structural continuity. The all-markers model successfully reconstructs a wide range of morphologies, demonstrating good generalization. Lower mask ratios (e.g., 0.15) help preserve spatial context for dense cell types, whereas moderate ratios (e.g., 0.55) provide more effective regularization for sparsely distributed markers.

6.4. CANVAS dataset structure

Each cell type contains a full brain volume as sequential TIFF Z-slices and ground truth annotations for six ROIs split into train and test sets, while test sets are not publicly announced. The compressed .tar files are also available for each dataset.

```
CANVAS
└── [Marker]_[BrainID] (e.g., IBA1_brain_11)
    ├── [Marker]_[BrainID].tar
    ├── image
    │   ├── 344170_392340_000000_ch2.tiff
    │   ├── 344170_392340_000010_ch2.tiff
    │   └── ... (sequential Z-slices)
    ├── label
    │   ├── TRAIN
    │   │   ├── [Marker]_[brain]_region1_gt.csv
    │   │   ├── [Marker]_[brain]_region2_gt.csv
    │   │   └── [Marker]_[brain]_region3_gt.csv
    │   └── TEST (not publicly available)
    └── LICENSE.md
```

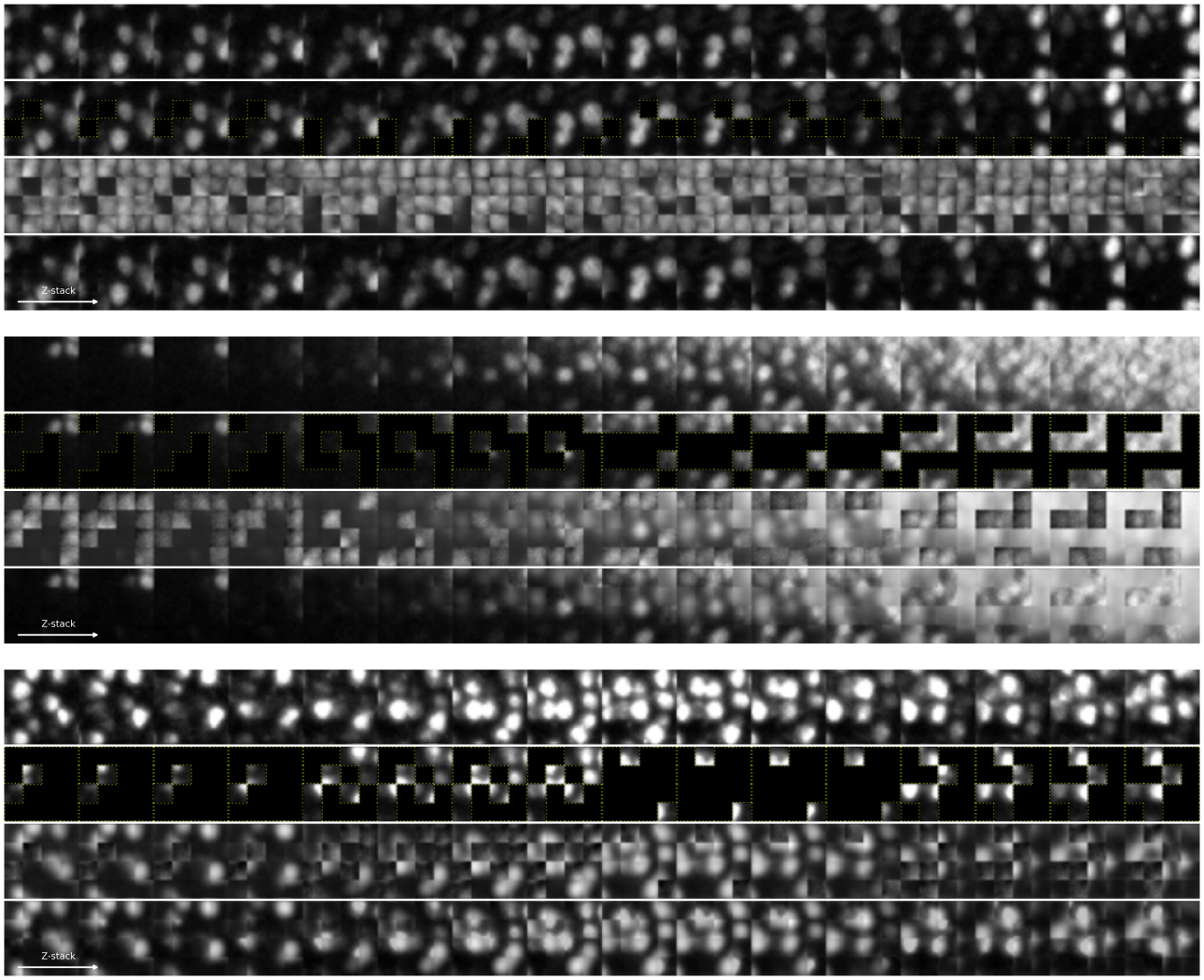

Figure 5. **NeuN reconstruction with varying mask ratios.** Compact neuronal nuclei reconstructed with 16×32×32 crop, 4×8×8 patch. Top to bottom: mask ratio 0.15, 0.55, 0.75. Lower masking preserves more morphological detail.

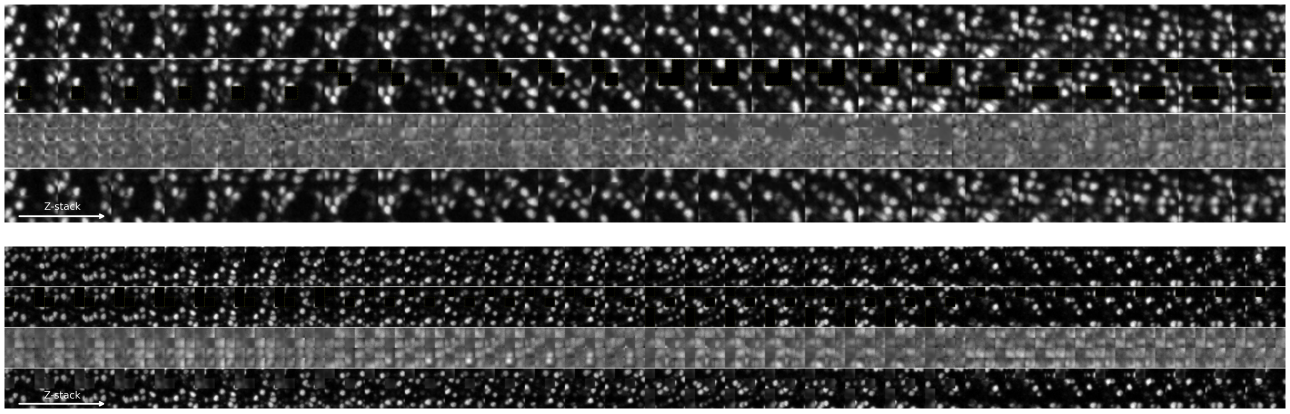

Figure 6. **NeuN reconstruction with varying crop sizes.** Mask ratio 0.15. Top: 24×48×48 crop (6×12×12 patch). Bottom: 32×64×64 crop (8×16×16 patch). Larger crops capture more spatial context.

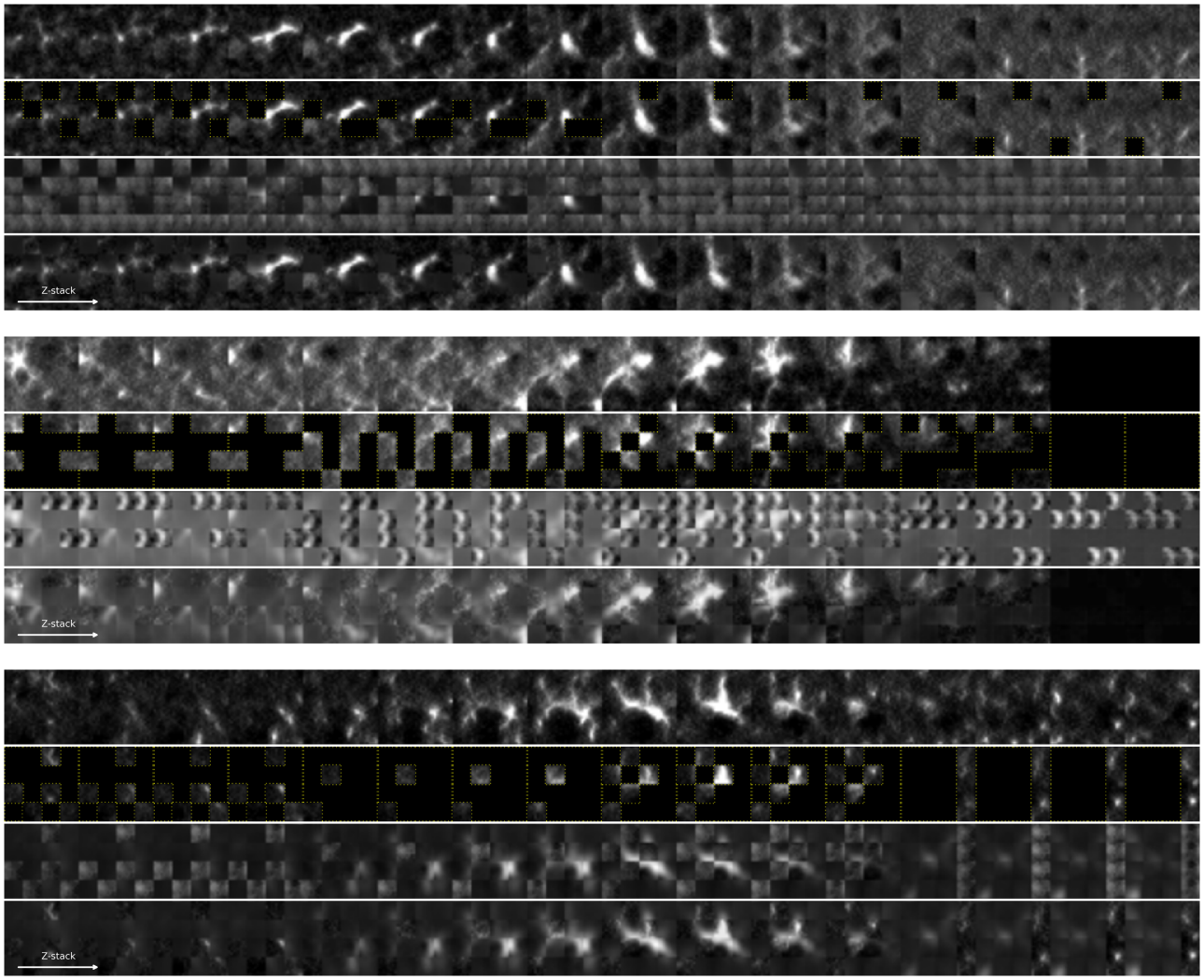

Figure 7. **IBA1 reconstruction with varying mask ratios.** Ramified microglial morphology with fine processes. 16×32×32 crop, 4×8×8 patch.

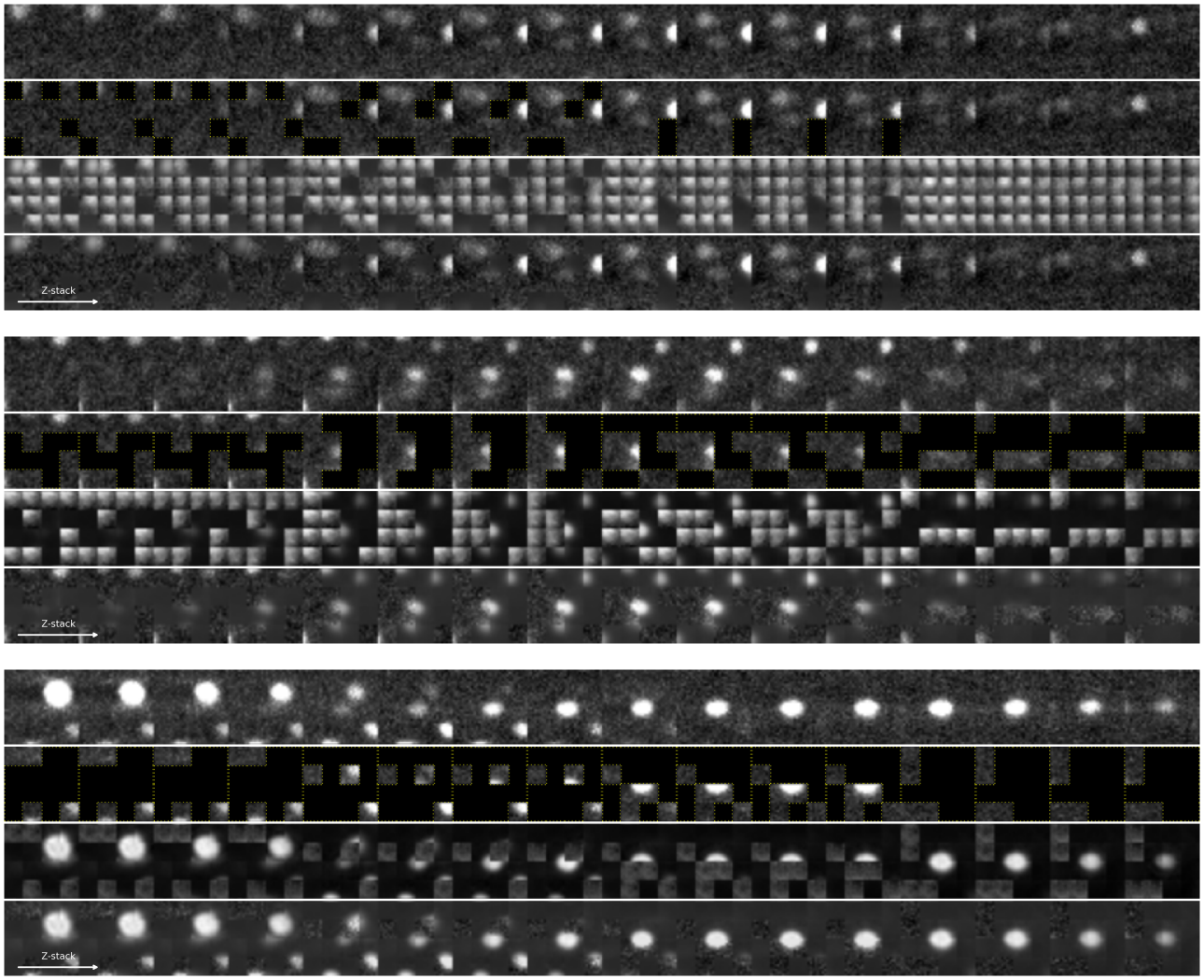

Figure 8. **cFos reconstruction with varying mask ratios.** Sparse activity-dependent signal. 16×32×32 crop, 4×8×8 patch.

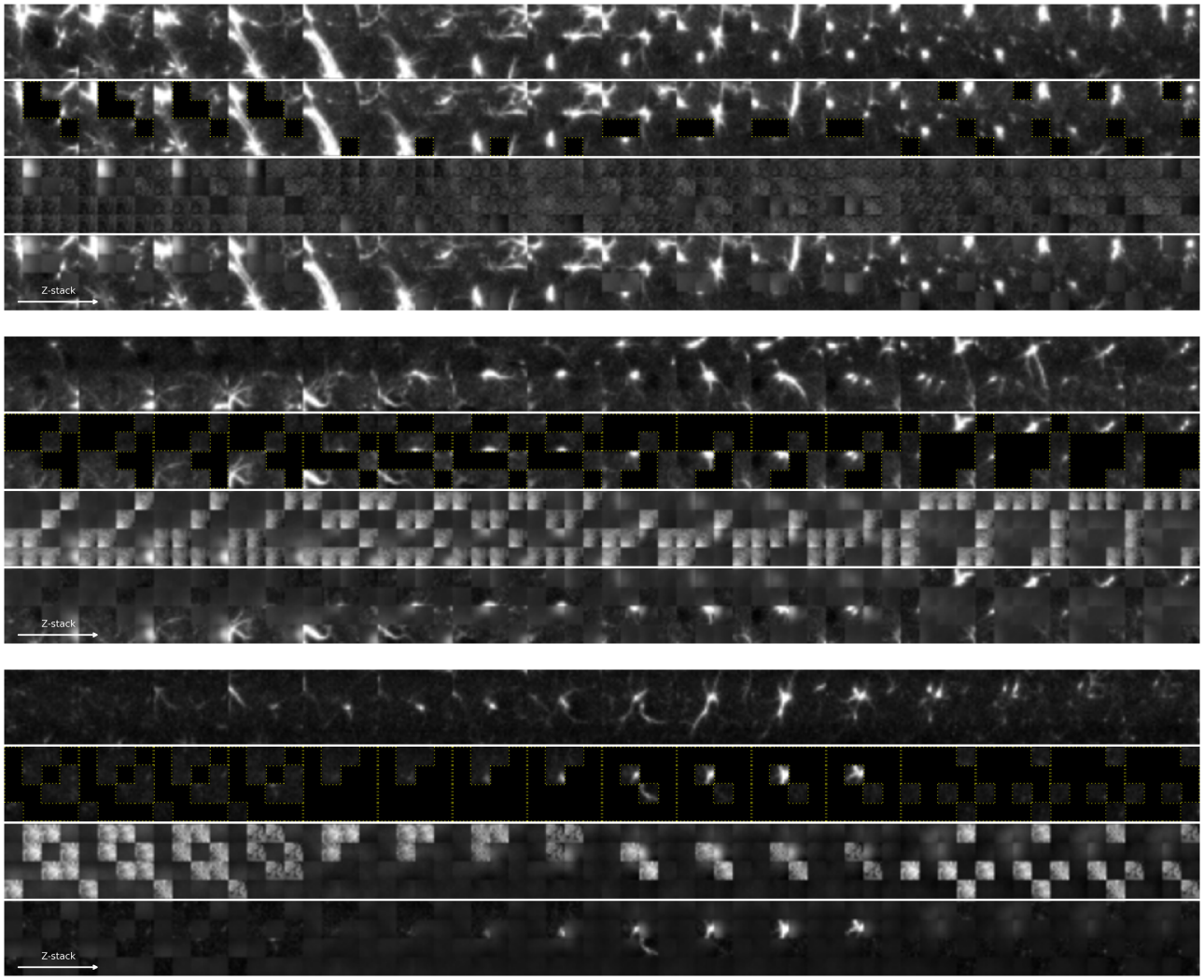

Figure 9. **GFAP reconstruction with varying mask ratios.** Complex astrocyte morphology with elongated processes. 16×32×32 crop, 4×8×8 patch.

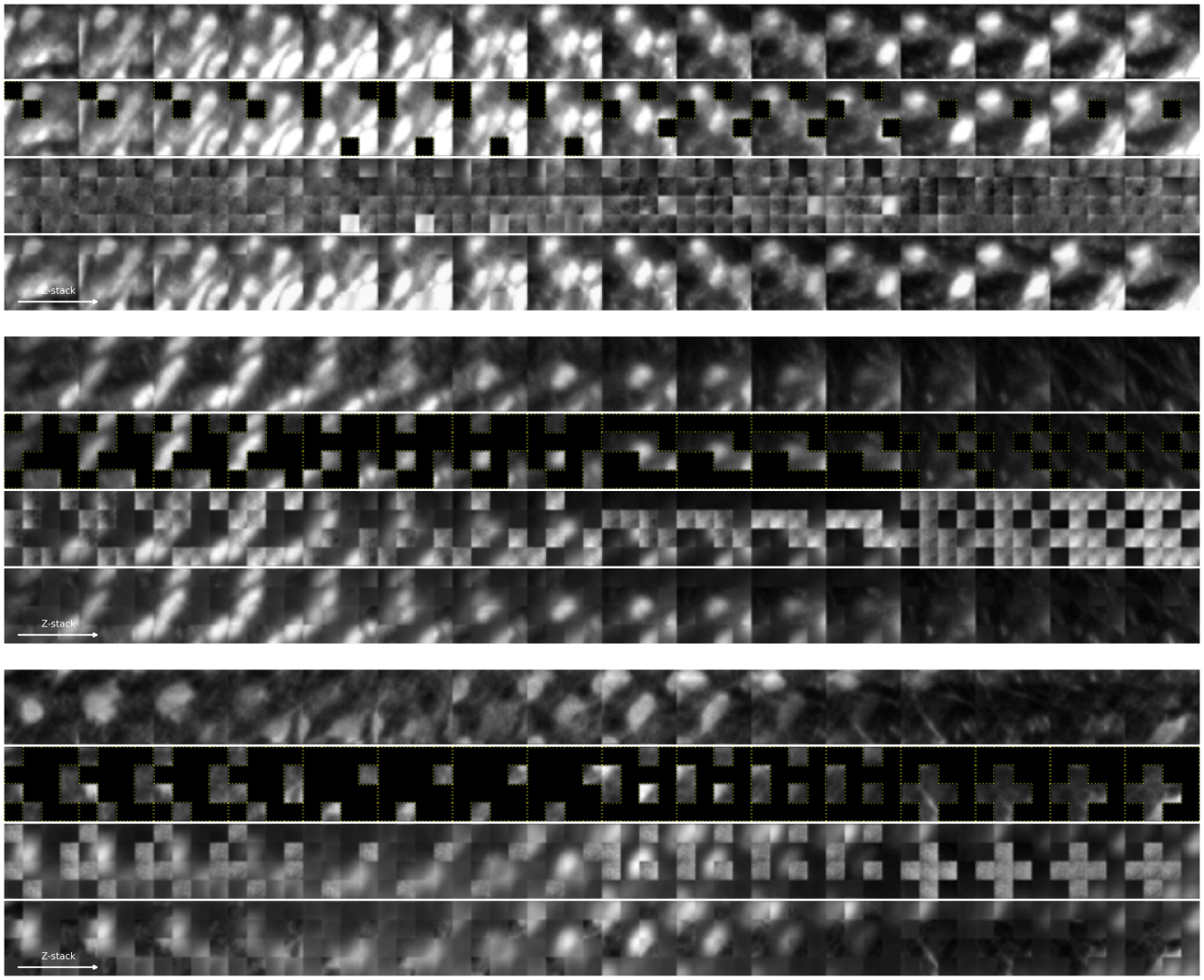

Figure 10. **PV reconstruction with varying mask ratios.** Sparse interneurons with distinct morphology. 16×32×32 crop, 4×8×8 patch.

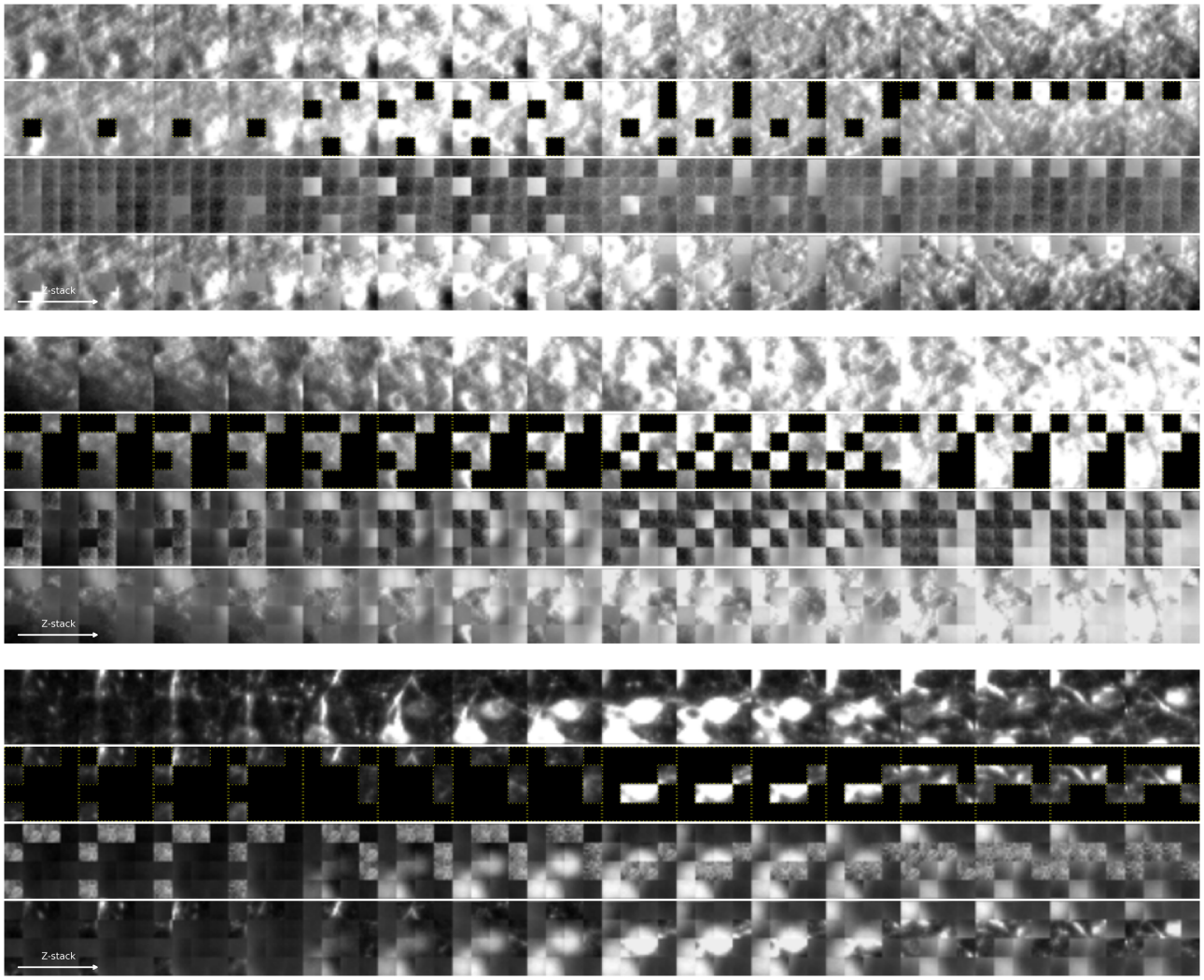

Figure 11. **TH reconstruction with varying mask ratios.** Elongated dopaminergic neuron morphology. 16×32×32 crop, 4×8×8 patch.

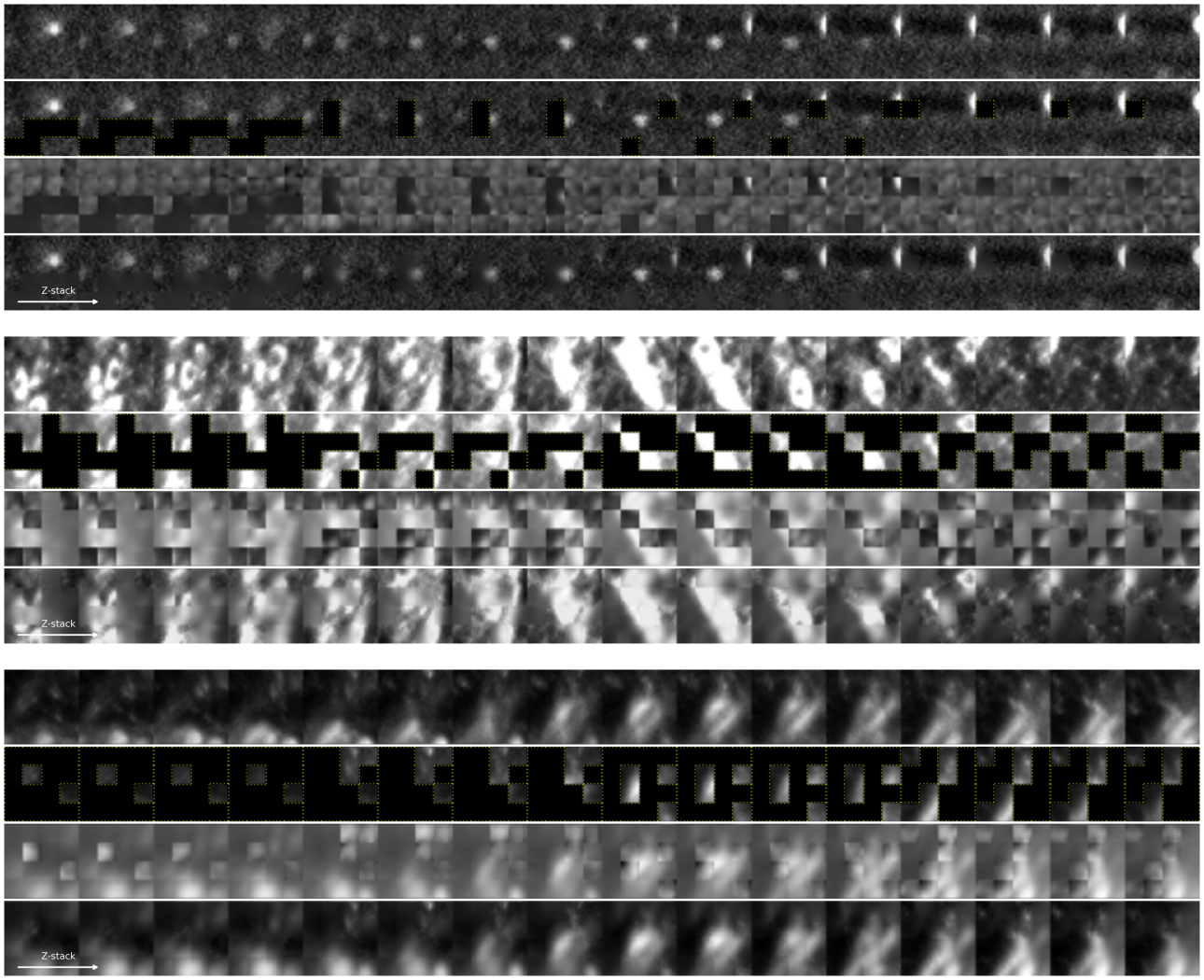

Figure 12. **All-markers universal model reconstructions.** Training on diverse cell types (60k patches from all 6 markers) produces a generalizable encoder maintaining quality across morphologies. 16×32×32 crop, 4×8×8 patch.