Robust virtual staining of landmark organelles

Ziwen Liu[1*], Eduardo Hirata-Miyasaki[1*], Soorya Pradeep[1], Johanna Rahm[1,2], Christian Foley[1,3], Talon Chandler[1], Ivan Ivanov[1], Hunter Woosley[1], Tiger Lao[1], Akilandeswari Balasubramanian[1], Chad Liu[1], Manu Leonetti[1], Carolina Arias[1], Adrian Jacobo[1], Shalin B. Mehta[†]

[1]Chan Zuckerberg Biohub San Francisco, San Francisco, USA; [2]Institute of Physical and Theoretical Chemistry, Goethe University, Frankfurt, Germany; [3]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, USA

* These authors contributed equally to this work.

[†]correspondence: shalin.mehta@czbiohub.org

Abstract

Dynamic imaging of landmark organelles, such as nuclei, cell membrane, nuclear envelope, and lipid droplets enables image-based phenotyping of functional states of cells. Multispectral fluorescent imaging of landmark organelles requires labor-intensive labeling, limits throughput, and compromises cell health. Virtual staining of label-free images with deep neural networks is an emerging solution for this problem. Multiplexed imaging of cellular landmarks from scattered light and subsequent demultiplexing with virtual staining saves the light spectrum for imaging additional molecular reporters, photomanipulation, or other tasks. Published approaches for virtual staining of landmark organelles are fragile in the presence of nuisance variations in imaging, culture conditions, and cell types. This paper reports model training protocols for virtual staining of nuclei and membranes robust to label-free imaging parameters, cell states, and cell types. We developed a flexible and scalable convolutional architecture, named UNeXt2, for supervised training and self-supervised pre-training. The strategies we report here enable robust virtual staining of nuclei and cell membranes in multiple cell types, including neuromasts of zebrafish, across a range of imaging conditions. We assess the models by comparing the intensity, segmentations, and application-specific measurements obtained from virtually stained and experimentally stained nuclei and membranes. The models rescue the missing label, non-uniform expression of labels, and photobleaching. We share three pre-trained models, named VSCyto3D, VSCyto2D, and VSNeuromast, as well as VisCy, a PyTorch-based pipeline for training, inference, and deployment that leverages the modern OME-Zarr format.

Introduction

Building predictive models of complex biological systems requires technologies to visualize and model the interactions among cells and organelles. Multiplexed dynamic imaging of organelles and cells is limited by trade-offs between spatial resolution, temporal resolution, number of channels, and photodamage. These trade-offs are compounded in high-throughput experiments that incorporate diverse perturbations and cell types. We illustrate these trade-offs with two problem spaces: A) Image-based phenotyping of the cell dynamics at single-cell resolution (1–4) requires multiplexed imaging of organelles, nuclei, and cell membranes over time and across perturbations. In this case, the multispectral fluorescent imaging of organelles, nuclei, and cell membranes limits the throughput. B) Understanding the mechanisms of emergence and homeostasis of cell types during the development of an organ, such as the zebrafish neuromast (5, 6), requires tracking individual cell types and developmental signals. Multiplexing fluorescent reporters of developmental signals, cell type, nuclei, and membrane without perturbing the development is challenging and engineering embryos with multiple reporters is labor-intensive. In both cases, the phototoxicity and photobleaching induced by imaging multiple channels limits the duration of experiments.

Correlative label-free and fluorescence imaging, combined with demultiplexing of cellular components with deep learning are emerging as a promising and widely useful solution to the problem of multiplex dynamic imaging and analysis (7–11). 3D quantitative phase imaging methods (8, 9, 12–15) encode the dry mass of several “landmark organelles” such as nuclei, cell membrane, nucleoli, nuclear envelope, and lipid droplets in a single channel with high accuracy. Quantitative polarization imaging methods encode alignment and orientation of ordered organelles such as cytoskeleton and can be multiplexed with phase imaging (8, 12, 16). Virtual staining is an image translation task that transforms the measurements of these physical properties into molecular labels of organelles. Virtual staining of quantitative label-free images can enable sensitive analysis of the organelle interactions without the need for laborious and error-prone human annotations of voxels in movies. In addition to the image-based phenotyping of cell dynamics, image translation is used for cross-modal image registration (7, 9), separating organelles labeled with the same fluorophore (17), and rapid 3D histology (18, 19).

Published work on virtual staining of label-free images and separating organelles imaged with the same fluorophore implies that virtual staining can indeed relax the longstanding multiplexing bottleneck in dynamic imaging. Then, why is this approach not used more widely? The key bottleneck(7, 8, 11) is that virtual staining models do not generalize to imaging parameters, cell states, and cell types beyond the ones included in training data. In this paper, we tackle this challenge by reporting strategies to train generalist virtual staining models of nuclei and membranes that make them insensitive to the changes in imaging parameters irrelevant for image-based screens and sensitive to relevant parameters. More precisely, we train the models whose outputs are invariant relative to the nuisance changes in label-free input (e.g., noise, numerical aperture, phase contrast modality) and equivariant to important changes in label-free input (e.g.,

magnification). We also report training protocols for few-shot generalization of the models to unseen developmental stages and cell types.

Segmentation of nuclei and cells is a common first step in image-based phenotyping, including in the questions we mentioned above. Several generalist nuclei and cell membrane segmentation models have been reported with diverse architectures (20–23) and are used for the segmentation of nuclei and cells from fluorescence images. The majority of these models were not trained with high-resolution quantitative label-free imaging methods and do not generalize to these datasets. Fine-tuning these models for label-free images requires expensive human annotation. We show that the combination of generalist virtual staining with published generalist fluorescence segmentation models enables reliable single-cell analysis.

As virtual staining or image translation models have been developed, a variety of convolution and attention-based architectures have been explored. There is an active debate whether transformer models that use attention operation (21–24) fundamentally outperform convolutional neural networks that rely on the inductive bias of shift equivariance. Systematic comparisons suggest that convolutional models perform as well as transformer models (24, 25) when large compute budget is spent, and outperform the transformer models when moderate compute budget is spent. In this study, we use a *purely convolutional* architecture that draws on the design principles of transformer models. Our choice of architecture is inspired by U-Net (26), ConvNeXt v2 (27, 28) and SparK (29).

This paper makes the following specific contributions: a) data augmentation strategies inspired by the physics of image formation, b) training protocols that improve the generalization of the virtual staining models, c) UNeXt2, an efficient image translation architecture inspired by U-Net, ConvNeXt v2, and SparK, d) trained models for virtual staining of nuclei and membrane from widely deployable zernike phase contrast or quantitative phase contrast data, and e) a permissively licensed pythonic pipeline, named VisCy (30), for model training, inference, and evaluation that implements strategies reported here and uses a modern image format standard, OME-Zarr (31, 32), as input. We assess the gains in performance due to architectural refinement, augmentation strategies, and training protocols using a suite of metrics that include regression metrics, instance segmentation metrics, and application-specific metrics relevant to the two questions mentioned above.

The results are organized as follows: We first summarize the model architecture and metrics for robust virtual staining. Second, we demonstrate the model's invariance to nuisance imaging parameters and equivariance to magnification in a specific cell type. We then show a fine-tuning strategy for long-term imaging experiments typical of developmental biology. Last, we demonstrate a pre-training/fine-tuning paradigm for generalizing virtual staining to new cell types.

Results

Models and metrics for robust virtual staining

Virtual staining models are trained using paired label-free and fluorescence images (Figure 1A, orange and blue arrows). Once a model is trained, only the label-free input is needed for inference (Figure 1A, orange arrows). This paper presents models trained with an inexpensive quantitative phase imaging modality (phase from defocus) as their input, which can be implemented on any motorized widefield microscope. This straightforward computational imaging technique involves acquiring a z-stack in brightfield and reconstructing phase density using an image formation model that relates the 3D intensity distribution with the specimen's scattering potential (8, 33, 34).

Design choices from UNet (26), convNeXt v2 (28), and SparK (29) architectures were integrated to develop an efficient and scalable architecture, named UNeXt2 (Figure 1A). UNeXt2 implements a projection module in the stem and head of the network that provides a flexible choice of the number of slices in the input stacks and output stacks, enabling 2D, 2.5D, and 3D image translation(8) with the same architecture. The body of the network is a UNet-like hierarchical encoder and decoder with skip connections that learns a high-resolution mapping between input and output. The UNeXt2 architecture provides 15x more learnable parameters for 3D image translation than our previously published 2.5D UNet at the same computational cost (Table 1). The efficiency gains are even more significant when compared to 3D UNet. This approach enables the allocation of the available computing budget to train moderate-sized models faster or to train more expressive models that generalize to new imaging conditions and cell types. The models trained for joint prediction of nuclei and membrane are slightly more accurate than models trained for prediction of nuclei alone (compare metrics shown in Table 1). The choice of layers, blocks, and the loss function are described in Methods and Table 2.

This paper focuses on three models for joint virtual staining of nuclei and membrane (Figure 1B) that address distinct use cases: analyzing cell states in HEK293T cells from the OpenCell library (VSCyto3D), 3D virtual staining of neuromasts for analyzing cell growth and death during development (VSNeuromast), and 2D virtual staining for high throughput screens across multiple cell types, HEK293T, A549, BJ-5ta (VSCyto2D). In all of these applications, virtual staining and generalist segmentation models are used in tandem to segment the nuclei and cells from label-free images. The phase measurements and complementary fluorescence sensors are then used for quantitative analysis of functional states of cells with single-cell resolution (Figure 1A, single-cell phenotyping).

The experimentally and virtually stained nuclei and cell membrane are segmented with Cellpose (see Methods, Figure S1). The Cellpose model requires significant fine-tuning with label-free images but works well with virtually stained images of nuclei and cell membrane, primarily because the training set of Cellpose included only classical Zernike phase contrast (35) and fluorescence data. As can be seen from images in Figure 1B, virtually stained images are intrinsically denoised because the models

cannot learn to predict random noise. This feature obviates the need to train models that are robust to noise, such as Cellpose3 (21). Joint virtual staining of nuclei and membranes enables more accurate cell segmentation (20).

The performance of the virtual staining models is assessed with the regression metrics (Pearson correlation coefficient, PCC) and instance segmentation metrics (average precision, AP) between experimentally stained images and virtually stained images. Due to the variations in experimental labeling and the need to fine-tune Cellpose models to new cell shapes, we cannot rely on experimental fluorescence images and their segmentations obtained with Cellpose as absolute ground truth. For example, it is challenging to segment boundaries of BJ-5ta cells at low magnifications (Figure 1B, Figure S1), because they have diverse shapes and grow on top of each other. Therefore, this paper takes the approach of first comparing the experimental and virtually stained images and their segmentations, and then, quantifying the observations with metrics. The model refinement and hyperparameter optimization are guided by application-driven metrics such as cell size of cultured cells and nuclei count in neuromasts (Figure 1C), in addition to regression and segmentation metrics.

VSCyto3D and VSNeuromast are trained in a supervised fashion using UNeXt2 architecture and mixed loss. VSNeuromast is fine-tuned to new developmental stages using sparsely sampled time series of label-free and fluorescence volumes. VSCyto2D model is trained with a pre-training/fine-tuning paradigm common for language and vision transformers. Subsequent results describe each of these training protocols and our findings on the regime of generalization of the resulting models.

Virtual staining robust to imaging parameters

Nuisance variations in label-free images often degrade the performance of virtual staining models. In this section, we explore preprocessing, data augmentation, and sampling strategies to train virtual staining models robust to variations in the label-free contrast, resulting in the VSCyto3D model.

Deconvolution of raw intensities is an effective preprocessing step to remove some of the nuisance variations in the contrast, such as non-uniform illumination, and to improve contrast of biological structures in the image data. As shown in Figure 2A, deconvolution of phase density (Phase) from brightfield (BF) data (8, 33) and deconvolution of fluorescence density (FL density) from raw fluorescence (FL) improves the contrast for organelles, which are typically encoded in the middle spatial frequencies of the passband of the microscope. The restoration of phase density from raw intensities enable quantitative interpretation of the dry mass of the cells. In brightfield images, dense structures are transparent in focus, and brighter or darker relative to the background when out of focus. In the deconvolved phase density images, the contrast is more uniform (Figure 2A). Deconvolution also removes slowly varying intensity and phase variations that typically arise due to non-uniform illumination or meniscus of fluid that forms in imaging chambers.

Four virtual staining models that translate to and from combinations of raw and deconvolved images were trained using UNeXt2 architecture (Figure S2) as shown in

Figure 2A. The model trained to predict fluorescence density from phase density leads to the sharpest predictions of nuclei and membrane. This is reflected in higher average precision (AP) and higher AP at the IoU of 0.5 (AP@0.5) between the instance segmentations of nuclei obtained from fluorescence density and predictions of virtual staining models (Figure S3). Interestingly, the deconvolution causes a drop in Pearson correlation coefficient (PCC) between virtually stained and fluorescence density images, because the virtually stained fluorescence is inherently smooth (Figure S3). The smoothing enables robust segmentation of landmark organelles in the presence of noise, but is detrimental for virtual staining of structures close to the resolution limit of the microscope. The improvement in segmentation metric, but worsening of regression metric due to deconvolution of data illustrates the need for nuanced interpretation of the metrics.

Data augmentations that account for the formation of natural and medical images have been important for robust representation learning (36) and segmentation (37). We reasoned that training data should be augmented with spatial and intensity filters inspired by the image formation of microscopes to make the predictions of our models invariant to nuisance variations in imaging conditions, e.g., exposure, noise, and the size of the illumination aperture. Figure 2B illustrates the images without and with such spatial and intensity augmentations (Methods). The predictions (Figure 2B, virtual staining with augmnetations) and segmentations (Figure S4) across the test dataset become invariant to variations in imaging parameters as we incorporate spatial and intensity augmentations inspired by image formation. As expected, the scaling augmentations make the model equivariant to magnification.

Fluorescent labeling is stochastic, especially when cells are engineered to express multiple fluorescent tags (38). Sampling the patches from the training data in proportion to the degree of labeling makes the models robust to partial and uneven labeling as shown in Figure 2B (white box). In fact, the VSCyto3D model rescued the nuclear stain in the fields of view from the test dataset (Figure S5) where many cells were missing the nuclear stain. Comparison of the 3D distribution of experimentally and virtually stained nuclei and membrane in a through focus movie (Video 1) shows that virtual staining improves the uniformity of labeling of cell membrane.

We also explore if the label-free input images can be simulated to mimic acquisition with a different light path, which transfer less information than the light path with which the training data was acquired. Filters informed by image formation were included in the augmentation pipeline for input data to simulate the data acquisition with different light paths. This strategy enabled generalization of VSCyto3D model to Zernike phase contrast images (Figure 2C) not seen during the training. The raw fluorescence images of labeled nuclei and membranes were acquired with the Zernike phase contrast (PhC) objective were significantly blurrier and noisier (Figure 2C, raw fluorescence) than those acquired with the widefield objective, due to the phase ring specific to the PhC objective that absorbs and filters fluorescence emission. Interestingly, virtually stained nuclei and membranes are sharper (Figure 2C, virtual staining with augmentation), and therefore lead to sharper instance segmentations of nuclei and membranes (Figure S6). In other words, the augmentations we devised expanded the input image space to include

Zernike phase contrast-like images, while constraining the space of possible output images to sharp fluorescence density images. This strategy also enabled synthesis of training datasets at 20x magnification for training VSCyto2D model (Figure 1B, Methods).

The degree of perturbation to which the model is robust was assessed by simulating the blur and contrast stretch in the input image. The VSCyto3D model performs well, even when the images show significant blur and contrast variation (Figure S7). This implies that the model is robust to variations in numerical aperture that modulate the resolution and contrast of a phase image. In order to spot-check that the VSCyto3D model learns a meaningful mapping between imaging modalities, we visualize the feature maps (Methods - Model visualization) at each level of encoder and decoder (Figure S8) for a randomly chosen test image. The boundaries of cells and nuclei can be identified at higher levels of abstraction in the encoder and the decoder.

These results demonstrate a training protocol for robust virtual staining that consists of acquiring training data at the highest possible resolution, deconvolving it with an image formation model, augmenting with filters to mimic the changes in contrast and resolution, and sampling it in proportion to the degree of labeling.

Virtual staining in developing 3D organs

The shape space of developing or differentiating cells evolves significantly and is often related to cell state and identity. For example, the cells that form the neuromasts of the zebrafish lateral line exhibit complex three-dimensional shapes and textures that change throughout their development (5, 6). Using this organ as a model system, we explore an economical strategy to generalize 3D virtual staining models that can capture the complex architecture of the neuromast and the shape changes that arise during its development. Moreover, we show that this model can deal with the additional aberrations and imaging challenges characteristic of in-vivo time-lapse experiments.

VSNeuromast model is trained with UNeXt2 architecture with 21 z-slices as input and output (Figure 1, Figure S9) following the training protocol described for VSCyto3D model. The model is first trained with data acquired on the widefield fluorescence microscope at two developmental stages (3dpf and 6.5dpf, dpf = days post fertilization).

When the model was used to predict nuclei and membrane at 4 dpf and compared with confocal images of nuclei and membrane, no hallucinations were noticed, but the predictions were blurry (Figure 3A, row: virtual staining without fine-tuning). Images at 4dpf are slightly out of the distribution of the training data acquired at 3dpf and 6.5dpf.

An economical strategy for generalization of virtual staining across developmental time without the risk of hallucinations is to fine-tune a pre-trained model with coarsely sampled pairs of label-free and fluorescence data. With this strategy, nuclei and cells can be tracked at high temporal resolution with phase imaging and the model's accuracy can be continuously calibrated. This approach can reduce the photodamage, enabling faster and longer imaging of developmental dynamics. To assess the viability of this approach, the movie of neuromasts imaged at 4dpf is subsampled by the factor

of 1/12 to create a fine-tuning training set. The fine-tuning workflow is described in Methods and Figure S10. Visualization of predictions on a neuromast from the validation set (Video 2) illustrates how blurry predictions of the pre-trained model are sharpened during fine-tuning.

The fine-tuned model is more accurate (Figure 3A, row: fine-tuned virtual staining). As shown in Figure 3B, fine-tuning improves the PCC between the virtually stained nuclei and membrane and experimentally stained nuclei and membrane. The fine-tuned model enables robust virtual staining of nuclei and membrane in 3D and over time as seen from the comparison of experimental and virtually stained neuromast in Video 3.

The virtually stained nuclei show a more uniform intensity distribution than the experimentally stained nuclei as seen from Figure 3A and Figure 3C. The dimmer nuclei in the fluorescence density image are missed by the segmentation model but are rescued by virtual staining (Video 4). A comparison of the mean cell count over time obtained from fluorescently labeled and virtually stained nuclei of five lateral line neuromasts corroborates the rescue (Figure 3B, cell count, green curve vs blue curve) of dim nuclei seen in Video 4 and Figure 3C. A comparison of instance segmentation from experimentally and virtually stained membrane shows less pronounced, yet measurable, rescue of cells by virtual staining (Figure 3B, cell count, magenta curve vs orange curve). In both experimental and virtually stained neuromasts, we notice extraneous cell segmentations at the edge of neuromasts. We anticipate filtering these extraneous segmentations via tracking.

The mean intensity of each segmented cell's nuclei and membrane were compared (Figure 3D) to assess whether the model can rescue photobleaching. The experimental membrane channel shows clear photobleaching. On the other hand, the photobleaching is measurably corrected by the virtual staining model. Thus, substituting an experimental stain with a virtual stain can reduce photobleaching/photodamage.

When the VSNeuromast model was used for inference across zebrafish, it led to the hallucination of cells around the yolk, because the size and texture of these cells resemble cells of neuromast. However, these cells can be easily filtered in post-processing as shown in Figure S11.

The feature maps learned by the encoder and decoder of the VSNeuromast model were visualized to assess the transformation learned by the model (Methods). The model indeed represents shapes of nuclei, membrane, and neuromast as seen from the principal components of the feature maps shown in Figure S12 for an example input image of neuromast.

Self-supervised pre-training for few-shot generalization

The results above demonstrate training protocols for generalizing the virtual staining models for a given cell type to new imaging conditions and in a developing organ. Here, we discuss generalization of virtual staining models to diverse cell types of interest in image-based drug screens.

Virtual staining can accelerate single-cell phenotyping with diverse cell types, but the need to collect paired training data of sufficient diversity is often challenging. For example, consistent labeling of cell membrane is currently practical only with genetically expressed peptides (e.g., CAAX) that localize to cell membrane. Engineering the cells to label cell membranes and other landmark organelles is easy in transformed cell lines, but challenging in sensitive cell lines that more accurately represent human biology. Since the shapes of several landmark organelles are consistent across cell types, we reasoned that employing the pre-training/fine-tuning paradigm that is common in language and vision modeling can enable few-shot generalization to a new cell type.

Figure 4A-C illustrates a self-supervised pre-training and supervised fine-tuning protocol with one cell type (HEK293T). The phase images are randomly masked, and the unmasked pixels are used to predict the masked pixels in each training patch (Methods, Figure 4A), following the fully convolutional masked autoencoder (FCMAE) protocol reported for image classification (28). The computational graphs of the models used for pre-training and fine-tuning are shown in Figure S13. The virtually stained images in Figure 4B show that the pre-training/fine-tuning protocol slightly improves the visual sharpness of predicted images. As shown earlier (Figure S5), some of the fields of view had missing nuclei labels. These fields of view were proofread and segmentation and regression metrics (Figure 4C) were computed from virtually stained and experimental fluorescence images. These metrics show that the models pre-trained on label-free images with masked autoencoder, and fine-tuned on paired data are as accurate as models trained from scratch.

Figure 4D illustrates a pre-training/fine-tuning protocol for few-shot generalization to a new cell type. The model is pre-trained in two steps: 1) The encoder and decoder weights are optimized with just phase images of HEK293T and A549 cells using the masked autoencoding task shown in Figure 4A, 2) The weights are transferred to a virtual staining model that is pre-trained to predict fluorescent nuclei and membrane using HEK293T and A549 cells. After the pre-training, the model is fine-tuned with data acquired with a new cell type (BJ-5ta, fibroblast) that has a distinct morphology.

Video 5 shows that the model pre-trained with HEK293T and A549 datasets generalizes well to diverse cell shapes of A549 cells observed throughout the cell cycle. The images (Figure 4E) and segmentations (Figure S14) show that the pre-trained model after being fine-tuned with 6 FOVs performs as well as the model trained from scratch on with ~100 FOVs. Visualization of the evolution of the predictions from the validation set (Video 6) for the models trained with different training protocols are illustrated in Figure 4D, showing that pre-trained models produce correct predictions from the first epoch. Comparing the segmentation metrics for nuclei and membrane as a function of the number of training FOVs (Figure 4F) confirms that pre-trained/fine-tuned models scale better, i.e., generate more accurate predictions, relative to the models trained from scratch.

Finally, we visualize the feature maps of the models to assess the effect of different training protocols. We find that the model pre-trained on phase images (Figure S15 , column 3, rows: encoder stages) learns a more regular representation of cell

boundaries than the models trained on just the virtual staining task (Figure S15, columns 1 and 2, rows: encoder stages).

Taken together, the above results establish a training protocol for generalizing virtual staining models to new cell types.

Discussion and conclusion

The results of this study demonstrate new strategies for virtual staining of cellular landmarks. By integrating the design principles of UNet and ConvNeXt v2, the UNeXt2 model achieves high accuracy while reducing the computational cost relative to the earlier models, such as the 2.5D UNet. The UNeXt2 architecture's scalable and efficient design enables measurable improvement in virtually staining nuclei and membranes across diverse cell types and imaging conditions.

We report physics-informed preprocessing and data augmentations to improve the model's robustness to variations in imaging conditions, including unseen label-free imaging modality. These augmentations make the model robust to nuisance factors such as non-uniform illumination and changes in numerical aperture, ensuring the invariance of virtual staining results to these parameters. This robustness is particularly critical for practical image-based screens that integrate data from diverse microscopes with varying imaging conditions and optical aberrations.

The fine-tuning of models on subsampled time series, as illustrated with the VSNeuromast model, provides a compelling strategy for generalizing virtual staining across different developmental stages. This method reduces the risk of hallucinations and photodamage, enabling more accurate and extended imaging of dynamic cellular processes.

Finally, the paper reports a pre-training/fine-tuning protocol for few-shot generalization to new cell types. This method leverages the consistency of organelle shapes across different cell types, significantly reducing the data requirements for training robust virtual staining models. The performance improvement achieved by pre-training with a masked autoencoding task followed by fine-tuning suggests a viable method for scaling virtual staining models to a broader range of biological samples.

We provide a diverse set of evaluation metrics, including regression metrics, instance segmentation metrics, and application-specific measurements to comprehensively evaluate the models' performance. These metrics validate the accuracy and reliability of the virtual staining models, ensuring their applicability in real-world biological research.

Inspection of learned features confirms that the data augmentation strategies and training protocols described guide the model in learning a semantic mapping of cell structures between input and target imaging modalities. This capability is fundamental for the accurate virtual staining of cellular structures. We further illustrate failure modes and regime of validity of each of the three models reported here.

Future work will focus on several key areas. First, simulations with image formation models may further generalize the models to other phase imaging modalities without the need to acquire new data. Second, the pre-training/fine-tuning strategy can be extended to train decoders for other landmark organelles, such as nucleoli and lipid droplets. Third, the pre-training strategy can be extended across developmental stages, enabling the generalization of the models to evolving biological systems. Fourth, extending the feature analysis to statistical methods that rely on the whole test dataset can enable interpretation and guide the development of new training protocols, architectural refinements, and hyperparameter optimization. Finally, the training protocols developed for virtual staining can be adapted for segmentation models, potentially leading to joint virtual staining and segmentation models that offer even greater generalizability and accuracy.

In conclusion, this study presents a robust and scalable approach to virtual staining of nuclei and cell membranes, significantly improving their generalizability. The UNeXt2 architecture, combined with innovative data augmentation and pre-training strategies, enables accurate virtual staining across various cell types and imaging conditions. We hope that the release of these virtual staining models and the VisCy pipeline will facilitate the adoption and application of virtual staining techniques by the broader research community, thereby accelerating the mapping and understanding of dynamic cell systems.

Methods

Image acquisition and dataset curation

Human cell lines (HEK293T, A549, BJ-5ta)

HEK293T cells were labeled with H2B-mIFP and CAAX-mScarlet following the OpenCell protocol (39). Brightfield and fluorescence volumes of live HEK293T cells cultured in a 24-well plate. Training data were acquired on a wide-field fluorescence microscope (Leica Dmi8) with a 63x magnification, 1.3 NA glycerol-immersion objective. For testing robustness to different imaging conditions shown in Figure 2, additional volumes were imaged with a 40x magnification, 1.1 NA water-immersion objective, and a 100x magnification, 1.47 NA oil-immersion objective, at 0.25 µm Z-steps, for 96 Z-slices. For testing generalization to the Zernike phase contrast modality, image volumes were acquired with a 40x magnification, 0.6 NA Ph2 air objective, at 0.4 µm Z-steps, for 58 Z-slices. The images were sampled on a camera with 6.5 µm pixel size and 2048x2048 sensor For training, the Z-slice corresponding to the coverslip was determined by maximizing the transverse mid-band power of the mScarlet fluorescence density channel, and the Z-slices ranging from -2 µm to +12.5 µm relative to the coverslip were kept. Volumes that did not contain this range were excluded from the training dataset. The training dataset contains 291 volumes. For testing robustness to imaging conditions, images were acquired on a different day with new cell cultures, and 12 volumes were acquired for each condition.

A549 and BJ-5ta cells were stained with Hoechst for nuclei and CellMask for plasma membrane. Brightfield and fluorescence volumes of live cells cultured in 12-well plates were acquired on a wide-field fluorescence microscope (Leica Dmi8) with a 20x magnification, 0.55 NA objective. The images were sampled on a camera with 6.5 µm pixel size and 2048 by 2048 pixel sensor size, at 1 µm Z-steps. The A549 dataset was split into 24 volumes for training and validation, and 7 volumes for testing. The BJ-5ta dataset was split into 138 volumes for training and validation, and 12 volumes for testing.

Zebrafish neuromasts

Following the approved IACUC protocols, this study utilizes transgenic zebrafish lines expressing she:H2B-EGFP and cldnb:lyn-mScarlet (40) to label the nucleus and cell membrane, respectively, of the neuromasts. The VSNeuromast model utilizes neuromasts collected from three developmental stages: 3, 6 and 6.5 dpf (days post-fertilization: dpf). The datasets are composed primarily of lateral line neuromasts resulting in a dataset with 273 total volumes (160 volumes from 3dpf, 57 from 6dpf and 56 from 6.5dpf) with ZYX shape (107, 2048, 2048) or (35.3µm, 237.6µm, 237.6µm). The dataset was split 80/20 for training and validation respectively. Additionally, 6 neuromasts from an acquisition from a different fish and day are used as test dataset. For each neuromast, a brightfield and two fluorescence channel stacks were acquired at Nikon PlanApo VC x63 1.2NA objective on an ASI RAMM through the same optical path using a Andor ZYLA-4.2P-USB3-W-2V4. Channels were well registered because they shared the same imaging path. The fluorescence stacks of labeled nuclei and cell membrane were used as target channels.

The VSNeuromast model is fine-tuned for Figure 3 using the same transgenic line. Five lateral line neuromasts are imaged with an Olympus IX83 dual turret microscope using a 63x objective for fluorescence and label-free imaging every 5 minutes over a total of 12hrs. The VT-iSIM system with a Hamamatsu Quest v1 C15550-20UP with 4.6 µm pixel size is used for the fluorescence imaging. The label-free imaging uses a custom imaging path built on the first level of the microscope body using a 200mm tube lens (Thorlabs TTL-200A MP) resulting in 66x effective magnification using a machine vision camera (BFS-U3-51S5M-C). The test dataset was acquired with temporal interval of 5 minutes and the fine-tuning dataset is created by subsampling the timelapse by using the volumes acquired every hour. The fine-tuning datasets are split 80/20 into training and validation datasets.

Data conversion

All internal datasets are acquired in uncompressed lossless formats (i,e OME-TIFF and ND-TIFF) and converted to OME-Zarr using iohub (41), a unified python library to convert from most common bio-formats (i.e OME-TIFF, ND-TIFF, Micro-Manager TIFF sequences) and custom data formats to OME-Zarr.

Deconvolution of phase density and fluorescence density

The reconstruction from brightfield and fluorescence stacks to phase density and fluorescence density are reconstructed with the recOrder library (42) using the respective imaging parameters.

Registration

The label-free and fluorescence channels are registered with shrimPy (43). The resulting volumes are cropped to ZYX shape of (50, 2044, 2005) for the HEK293T Zernike phase contrast test dataset, (9, 2048, 2048) for A549, and (12, 2048, 2009) for BJ-5ta. The neuromast datasets acquired with the wide-field fluorescence microscope are registered to the phase density channel and cropped to (107, 1024, 1024). The datasets acquired in the iSIM setup are cropped to (81, 1024, 1024).

Model architecture

We use an asymmetric U-Net model with ConvNext v2 (28) blocks for both virtual staining and FCMAE pre-training. The original ConvNext v2 explored an asymmetric U-Net configuration for FCMAE pre-training and showed that it has identical fine-tuning performance on an image classification task. In the meantime, SparK (29) used ConvNext v1 blocks in the encoder and plain U-Net blocks in the decoder for its masked image modeling pre-training task. We use the 'Tiny' ConvNext v2 backbone in the encoder. For FCMAE pre-training, 1 ConvNext v2 block is employed per decoder stage. For virtual staining models, each decoder stage consists of 2 ConvNext v2 blocks.

Model training

Intensity statistics, including the mean, standard deviation, and median were calculated at the resolution of FOVs and at the resolution of whole dataset by subsampling each FOV using grid spacings of 32x32. These pre-computed metrics are then used to apply normalization transforms by subtracting the choice of median or mean and dividing by the interquartile range or standard deviation respectively. This enables z-scoring of the training data at the level of whole dataset, at the level of each FOV, and at the level of each patch (8), depending on the use case.

Training objectives

The mixed image reconstruction loss (44) is adapted as the training objective of virtual staining models: $\mathcal{L}^{\text{mix}} = 0.5 \cdot \mathcal{L}^{\text{2.5D MS-SSIM}} + 0.5 \cdot \mathcal{L}^{\ell_1}$. $\mathcal{L}^{\text{2.5D MS-SSIM}}$ is the multi-scale structural similarity index (45) measured without downsampling along the depth dimension, and $\mathcal{L}^{\ell_1}$ is the L1 distance (mean absolute error). The virtual staining performance of different loss functions is compared in Table 2.

The mean square error (MSE) loss is used to pre-train the FCMAE models on label-free images.

Data Augmentations

The data augmentations are performed with transformations from MONAI (37).

Spatial augmentation

MONAI transformation	Parameters
Random scaling	±0.3 or ±0.5 in XY, and ±0.2 in Z.
Random rotation	±π around the Z axis.
Random shearing	±0.05 in XY
Random XY flip	only used for VSCyto2D

Intensity augmentations

MONAI transformation	Parameters
Random contrast adjustment	gamma range (0.8, 1.2)
Random intensity scaling	±0.5
Random Gaussian blur:	sigma range (0.25, 0.75) in XY
Random Gaussian additive noise	strength determined per dataset (sigma 0.3-5).

VSCyto3D

Normalization

For each channel in the HEK293T dataset, the image volume is subtracted by its dataset level median and divided by the dataset level interquartile range.

Training

Models are trained with a warmup-cosine-annealing schedule on 4 GPUs with the distributed data parallel (DDP) strategy. A mini-batch size of 32 and learning rate of 0.0002 is used. Training and validation patch ZYX size is (5, 384, 384). For testing the effect of deconvolution (Figure 2B), models are trained for 100 epochs. For testing robustness to imaging conditions (Figure 2D), models are trained for 50 epochs.

VSCyto2D

Data pooling

Image volumes of HEK293T cells are downsampled from the 63x dataset with ZYX average pooling ratios of (9, 3, 3). For the VSCyto2D model reported in Figure 1,

training data are sampled from the downsampled HEK293T dataset, the A549 dataset, and the BJ-5ta dataset with equal weights.

Normalization

Each image volume is independently normalized before being used for model input. The phase channel is normalized to zero mean and unit standard deviation, and the fluorescence channels are normalized to zero median and unit interquartile range.

Pre-training

FCMAE pre-training with phase images is performed with a warmup-cosine-annealing schedule for 800 epochs on 4 GPUs with the DDP strategy and automatic mixed precision (AMP). A mini-batch size of 32 and learning rate of 0.0002 was used. Mask patch size is 32, and masking ratio is 0.5. Training and validation patch ZYX size is (5, 256, 256).

Fine-tuning

For fine-tuning, the encoder weights are loaded from FCMAE pre-trained models when applicable. The models are then trained for the virtual staining task with the encoder weights frozen or trainable. Models are trained on 4 GPUs with the DDP strategy and AMP. Training and validation patch ZYX size is (5, 256, 256). For testing data scaling with BJ-5ta, models are trained with a constant learning rate of 0.0002. 6-FOV models are trained for 6400 epochs, 27-FOV models are trained for 1600 epochs, and 117-FOV models are trained for 400 epochs.

VSNeuromast

Data pooling

The data used in our methods is pooled from four OME-Zarr stores, which contain neuromasts from 3dpf, 6dpf, and 6.5dpf stages. These stores include both the whole field of view (FOV) and a center-cropped version focused on the neuromast. For the cropped FOVs, a weighted cropping technique is applied to ensure the inclusion of training patches containing the neuromast. Conversely, the uncropped dataset employs an unweighted cropping method to incorporate additional contextual information. A high content screening (HCS) dataloader was developed to sample equally from the multiple datasets with variable length.

In the fine-tuning step, the experimental fluorescence channels were registered to the phase density channel and required downsampling of the data by the factor of 2.1 to match the pixel size between the datasets used for pre-training and fine-tuning.

Normalization

This model normalizes the label-free channel per FOV by subtracting the median and interquartile range.

Training

The models are trained with a warmup-cosine-annealing schedule on 4 GPUs with the distributed data-parallel (DDP) strategy. This model uses datasets from 3dpf and 6-6.5dpf using mini-batch size of 6 from each dataset and learning rate of 0.002.

The VSNeuromast model is initially trained using ZYX patch size of (15,384,384) for 150 epoch. The weights from this model are loaded to train a model that takes ZYX patch size of (21,384,384) to improve the Z prediction accuracy. Training and validation patch ZYX size is (21, 384, 384). This model is trained for 30 epochs.

Fine-tuning

The expression of nuclei and cell membrane labels in neuromast was equalized using a contrast adaptive histogram equalization (CLAHE) with a kernel size of [5,32,32] (z,y,x). The model is fine-tuned using the model's checkpoint ('neuromast_3n6dpf_21plane_v1_mixedloss_weightedcrop_hotstart_v1') using prior patch sizes (21,384,384), learning rate 2e-4 with a warmup-cosine-annealing schedule on 4 GPUs with DDP strategy. The model is trained with 45 epochs.

The fine-tuned model is used in all the neuromast figures. The fine-tune model generalizes to datasets using the same imaging setup the VSNeuromast model is trained on (Figure 1) and datasets acquired on iSIM setup described in Methods (Figure 3).

Inference using trained models

For the 2D virtual staining model VSCyto2D, each slice is predicted separately in a sliding window fashion.

For the 3D virtual staining models (ie. VSCyto3D and VSNeuromast), a z-sliding window is used. The predictions from the overlapping windows are then average-blended.

Model evaluation

The correspondence between fluorescence and virtually stained nuclei and plasma membrane channels are measured with regression and segmentation metrics. We describe the segmentation models for each use case below. All segmentation models are also shared with the release of our pipeline, VisCy (Code and Model Availability).

VSCyto2D

Segmentation of fluorescence density images as well as virtual staining prediction is performed with the 'nuclei' (nuclei) and 'cyto3' (cells) models in Cellpose. For BJ-5ta, a fine-tuned 'cyto3' model ('CP_20240530_060854') was used for cell segmentation. The nuclei segmentation target is corrected by a human annotator.

Pearson correlation coefficient (PCC) is computed between the virtual staining prediction and fluorescence density images. Average precision at IoU threshold of 0.5 (AP@0.5) is computed between segmentation masks generated from virtual staining images and segmentation masks generated from fluorescence density images.

VSCyto3D

Segmentation of H2B-mIFP fluorescence density and virtually stained nuclei is performed with a fine-tuned Cellpose 'nuclei' model ('CP_20220902_NuclFL'). The nuclei segmentation masks are corrected by a human annotator. Segmentation of cells

from CAAX-mScarlet fluorescence density and virtually stained plasma membrane is performed with the Cellpose ‘cyto3’ model. Due to loss of CAAX-mScarlet expression in some cells, positive phase density was blended with the CAAX-mScarlet fluorescence density to generate test segmentation targets. For the Zernike phase contrast test dataset, nuclei and cells are also segmented from the phase image using the Cellpose ‘nuclei’ and ‘cyto3’ models, in addition to segmentation from experimental fluorescence images.

PCC is computed between the virtual staining prediction and fluorescence density images. AP@0.5 and mean average precision of IoU thresholds from 0.5 to 0.95 at 0.05 interval (AP) is computed between segmentation masks generated from virtual staining images and segmentation masks generated from fluorescence density images.

VSNeuromast

The nuclei and cell segmentations of fluorescence images are generated with fine-tuned 3D Cellpose ‘nuclei’ model and from scratch using 19 manually corrected segmentations. The segmentations are human-corrected by using the napari-annotator plugin (https://github.com/bauerdavid/napari-nD-annotator) and morphological operators such as opening, closing, and dilation to remove artifacts. The nuclei segmentation model 'cellpose_Slices_decon_nuclei_nuclei_v7_2023_06_28_16_54' and the cell membrane segmentation model ‘'cellpose_2Chan_scratch_membrane_2024_04_01_17_12_00’ are used for neuromast segmentation across all figures.

The virtual staining models are evaluated by comparing the segmentations for fluorescence density and virtual staining predictions using Jaccard, Dice, and mean average precision (mAP) metrics at IoU thresholds of 0.5, 0.75, and 0.95. Additionally, PCC was computed between the prediction and the fluorescence density datasets.

Cellpose Segmentation Parameters

The following parameters are used for segmenting the experimental fluorescence and virtual staining to evaluate the respective models.

VSCyto2D

Parameters/Cellpose pre-trained model	**Nuclei Segmentation:** ‘nuclei’	**Cell body segmentation:** ‘cyto3’ (HEK293T and A549) / ‘CP_20240530_060854’ (BJ-5ta)
Diameter	0.0	0.0 / 150.0
Flow Threshold	0.4	0.4
Cell probability threshold	0.0	0.0
min_size	15	15

VSCyto3D

Parameters/Cellpose pre-trained model	**Nuclei Segmentation:** 'CP_20220902_NuclFL'	**Cell body segmentation:** 'cyto3'
Diameter	0.0	200.0
Flow Threshold	0.4	0.4
Cell probability threshold	0.0	0.4
min_size	15	15

VSNeuromast

Parameters/Cellpose segmentation model	**Nuclei Segmentation:** 'cellpose_Slices_decon_nuclei_nuclei_v7_2023_06_28_16_54`	**Cell body segmentation:** 'cellpose_2Chan_scratch_2024_04_30_11_12_00'
Diameter	60.0	65.0
Flow Threshold	0.0	0.4
Cell probability threshold	0.0	0.0
min_size	8000	8000

Model visualization

Principal component analysis of learned features

Each XY pixel in the output of a convolutional stage is treated as a sample with channel dimensions and decomposed into 8 principal components. The top-3 principal components are normalized individually and rendered as RGB values for visualization.

Code and Model Availability

The virtual staining pipeline is implemented as part of an open-source Python package for single-cell phenotyping, named VisCy (a contraction of words 'vision' and 'cell'). We use PyTorch (46) and PyTorch Lightning (47) as the training framework, MONAI (37) for data augmentation, and timm (48) for building blocks. Additionally, we implement custom I/O modules for reading and writing OME-Zarr stores during training and inference. The models referenced in the methods above are shared with releases of VisCy via GitHub (30).

Data Availability

Illustrative test datasets are accessible from scripts in released versions of VisCy. We will share the training datasets via public archive (such as Bioimage Archive) as the review of this work progresses.

Acknowledgments

Some of the computational experiments reported in this paper are informed by the discussions with participants of the advanced research course on deep learning at Marine Biological Lab (DL@MBL), Woods Hole. S. B. M. thanks fellow faculty Anna Kreshuk, EMBL, Heidelberg and Alex Krull, University of Birmingham, for critical discussion of strategies. E. H. thanks Ashesh, Human Technopole, Milan for pair coding image translation models during DL@MBL 2023. We thank Janie Byrum and George Bell, Chan Zuckerberg Biohub for their help with cell culture. We thank CZ Biohub's Scientific Computing team for enabling access to the high performance computing cluster. We thank Priscilla Chan and Mark Zuckerberg for supporting the CZ Biohub.

Author contributions

Z. L., E. H-M., S. P., J. R., I. I., H. W., T. L., A. B., C. L., A. J., and S. B. M. collected imaging data for training and testing the models. Z. L., E. H-M., S. P., C. F., and S. B. M. contributed to the development of the code for training models. Z. L. and E. H-M trained and evaluated models reported in this paper with guidance from S. B. M. Z. L., E. H-M, and S. B. M. wrote the paper with critical contributions from all co-authors.

Funding

All authors were supported by the intramural program of the Chan Zuckerberg Biohub, San Francisco. Johanna Rahm was supported by the DAAD (German Academic Exchange Service) IFI program (international research stays for computer scientists) and the DFG (German research foundation) iMOL (interfacing image analysis and molecular life-science) project number 414985841, GRK 2566.

This research was funded by the Chan Zuckerberg Biohub, San Francisco.

References

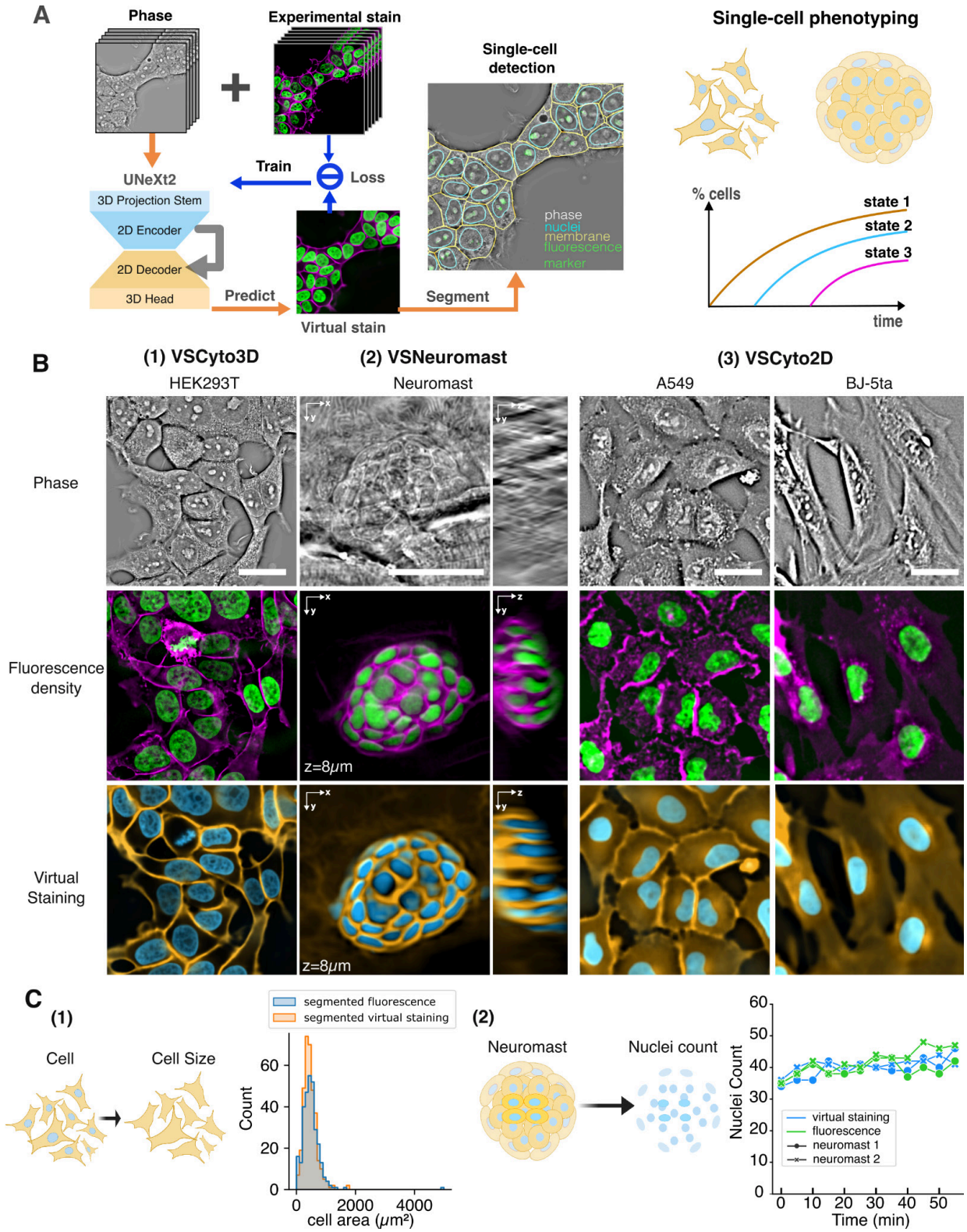
1. H. Kobayashi, K. C. Cheveralls, M. D. Leonetti, L. A. Royer, Self-supervised deep learning encodes high-resolution features of protein subcellular localization. *Nat. Methods* **19**, 995–1003 (2022).
2. M.-A. Bray, *et al.*, Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nat. Protoc.* **11**, 1757–1774 (2016).
3. Z. Wu, *et al.*, DynaMorph: self-supervised learning of morphodynamic states of live cells. *Mol. Biol. Cell* **33**, ar59 (2022).
4. J. Burgess, *et al.*, Orientation-invariant autoencoders learn robust representations for shape profiling of cells and organelles. *Nat. Commun.* **15**, 1022 (2024).
5. A. Jacobo, A. Dasgupta, A. Erzberger, K. Siletti, A. J. Hudspeth, Notch-Mediated Determination of Hair-Bundle Polarity in Mechanosensory Hair Cells of the Zebrafish Lateral Line. *Curr. Biol.* **29**, 3579-3587.e7 (2019).
6. M. N. Hewitt, I. A. Cruz, D. W. Raible, Data-Driven 3D Shape Analysis Reveals Cell Shape-Fate Relationships in Zebrafish Lateral Line Neuromasts. [Preprint] (2023). Available at: https://www.biorxiv.org/content/10.1101/2023.08.09.552694v1 [Accessed 26 May 2024].
7. C. Ounkomol, S. Seshamani, M. M. Maleckar, F. Collman, G. R. Johnson, Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. *Nat. Methods* **15**, 917 (2018).
8. S.-M. Guo, *et al.*, Revealing architectural order with quantitative label-free imaging and deep learning. *eLife* **9**, e55502 (2020).
9. I. E. Ivanov, *et al.*, Mantis: high-throughput 4D imaging and analysis of the molecular and physical architecture of cells. [Preprint] (2023). Available at: https://www.biorxiv.org/content/10.1101/2023.12.19.572435v1 [Accessed 5 January 2024].
10. J. Park, *et al.*, Artificial intelligence-enabled quantitative phase imaging methods for life sciences. *Nat. Methods* **20**, 1645–1660 (2023).
11. L. Kreiss, *et al.*, Digital staining in optical microscopy using deep learning - a review. *PhotoniX* **4**, 34 (2023).
12. I. E. Ivanov, *et al.*, Correlative imaging of the spatio-angular dynamics of biological systems with multimodal instant polarization microscope. *Biomed. Opt. Express* **13**, 3102–3119 (2022).
13. Y. Park, C. Depeursinge, G. Popescu, Quantitative phase imaging in biomedicine. *Nat. Photonics* **12**, 578 (2018).
14. R. Horstmeyer, J. Chung, X. Ou, G. Zheng, C. Yang, Diffraction tomography with Fourier ptychography. *Optica* **3**, 827–835 (2016).
15. O. Liba, *et al.*, Speckle-modulating optical coherence tomography in living mice and humans. *Nat. Commun.* **8**, 15845 (2017).
16. L.-H. Yeh, *et al.*, Permittivity tensor imaging: modular label-free imaging of 3D dry mass and 3D orientation at high resolution. *Nat. Methods* **In press**.
17. A. Ashesh, A. Krull, M. Di Sante, F. Pasqualini, F. Jug, uSplit: Image Decomposition for Fluorescence Microscopy in (2023), pp. 21219–21229.
18. Y. Winetraub, *et al.*, Noninvasive virtual biopsy using micro-registered optical coherence tomography (OCT) in human subjects. *Sci. Adv.* **10**, eadi5794 (2024).

19. B. Bai, *et al.*, Deep learning-enabled virtual histological staining of biological samples. *Light Sci. Appl.* **12**, 57 (2023).
20. M. Pachitariu, C. Stringer, Cellpose 2.0: how to train your own model. *Nat. Methods* **19**, 1634–1641 (2022).
21. C. Stringer, M. Pachitariu, Cellpose3: one-click image restoration for improved cellular segmentation. [Preprint] (2024). Available at: https://www.biorxiv.org/content/10.1101/2024.02.10.579780v2 [Accessed 7 April 2024].
22. A. Archit, *et al.*, Segment Anything for Microscopy. [Preprint] (2023). Available at: https://www.biorxiv.org/content/10.1101/2023.08.21.554208v1 [Accessed 7 April 2024].
23. J. Ma, *et al.*, The multimodality cell segmentation challenge: toward universal solutions. *Nat. Methods* 1–11 (2024). https://doi.org/10.1038/s41592-024-02233-6.
24. C. Stringer, M. Pachitariu, Transformers do not outperform Cellpose. [Preprint] (2024). Available at: https://www.biorxiv.org/content/10.1101/2024.04.06.587952v1 [Accessed 7 April 2024].
25. S. L. Smith, A. Brock, L. Berrada, S. De, ConvNets Match Vision Transformers at Scale. [Preprint] (2023). Available at: http://arxiv.org/abs/2310.16764 [Accessed 16 May 2024].
26. T. Falk, *et al.*, U-Net: deep learning for cell counting, detection, and morphometry. *Nat. Methods* **16**, 67–70 (2019).
27. Z. Liu, *et al.*, A ConvNet for the 2020s in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2022), pp. 11966–11976.
28. S. Woo, *et al.*, ConvNeXt V2: Co-Designing and Scaling ConvNets With Masked Autoencoders in (2023), pp. 16133–16142.
29. K. Tian, *et al.*, Designing BERT for Convolutional Networks: Sparse and Hierarchical Masked Modeling. [Preprint] (2023). Available at: http://arxiv.org/abs/2301.03580 [Accessed 26 May 2024].
30. Z. Liu, *et al.*, VisCy: computer vision models for single-cell phenotyping. (2023). Deposited 19 December 2023.
31. J. Moore, *et al.*, OME-NGFF: a next-generation file format for expanding bioimaging data-access strategies. *Nat. Methods* **18**, 1496–1498 (2021).
32. J. Moore, *et al.*, OME-Zarr: a cloud-optimized bioimaging file format with international community support. *Histochem. Cell Biol.* **160**, 223–251 (2023).
33. J. M. Soto, J. A. Rodrigo, T. Alieva, Label-free quantitative 3D tomographic imaging for partially coherent light microscopy. *Opt. Express* **25**, 15699–15712 (2017).
34. T. Chandler, L.-H. Yeh, I. Ivanov, C. Foltz, S. Mehta, waveorder. (2023). Deposited February 2023.
35. C. Edlund, *et al.*, LIVECell—A large-scale dataset for label-free live cell segmentation. *Nat. Methods* **18**, 1038–1045 (2021).
36. T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A Simple Framework for Contrastive Learning of Visual Representations in *Proceedings of the 37th International Conference on Machine Learning*, (PMLR, 2020), pp. 1597–1607.
37. M. J. Cardoso, *et al.*, MONAI: An open-source framework for deep learning in healthcare. (2022). https://doi.org/10.48550/arXiv.2211.02701.
38. A. M. Valm, *et al.*, Applying systems-level spectral imaging and analysis to reveal

the organelle interactome. *Nature* **546**, 162–167 (2017).
39. N. H. Cho, *et al.*, OpenCell: Endogenous tagging for the cartography of human cellular organization. *Science* **375**, eabi6983 (2022).
40. J. Peloggia, *et al.*, Adaptive cell invasion maintains lateral line organ homeostasis in response to environmental changes. *Dev. Cell* **56**, 1296-1312.e7 (2021).
41. Z. Liu, *et al.*, iohub. (2024). Deposited February 2024.
42. T. Chandler, *et al.*, recOrder. (2022). Deposited 23 August 2022.
43. I. E. Ivanov, E. Hirata-Miyasaki, T. Chandler, S. B. Mehta, czbiohub-sf/shrimPy. (2023). Deposited 19 December 2023.
44. H. Zhao, O. Gallo, I. Frosio, J. Kautz, Loss Functions for Neural Networks for Image Processing. [Preprint] (2018). Available at: http://arxiv.org/abs/1511.08861 [Accessed 30 August 2023].
45. Z. Wang, E. P. Simoncelli, A. C. Bovik, Multiscale structural similarity for image quality assessment in *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, (IEEE, 2003), pp. 1398–1402.
46. A. Paszke, *et al.*, PyTorch: An Imperative Style, High-Performance Deep Learning Library in *Advances in Neural Information Processing Systems 32*, H. Wallach, *et al.*, Eds. (Curran Associates, Inc., 2019), pp. 8024–8035.
47. W. Falcon, The PyTorch Lightning team, PyTorch Lightning. (2019). https://doi.org/10.5281/zenodo.3828935. Deposited March 2019.
48. huggingface/pytorch-image-models. (2024). Deposited 19 May 2024.

Figures

Figure 1



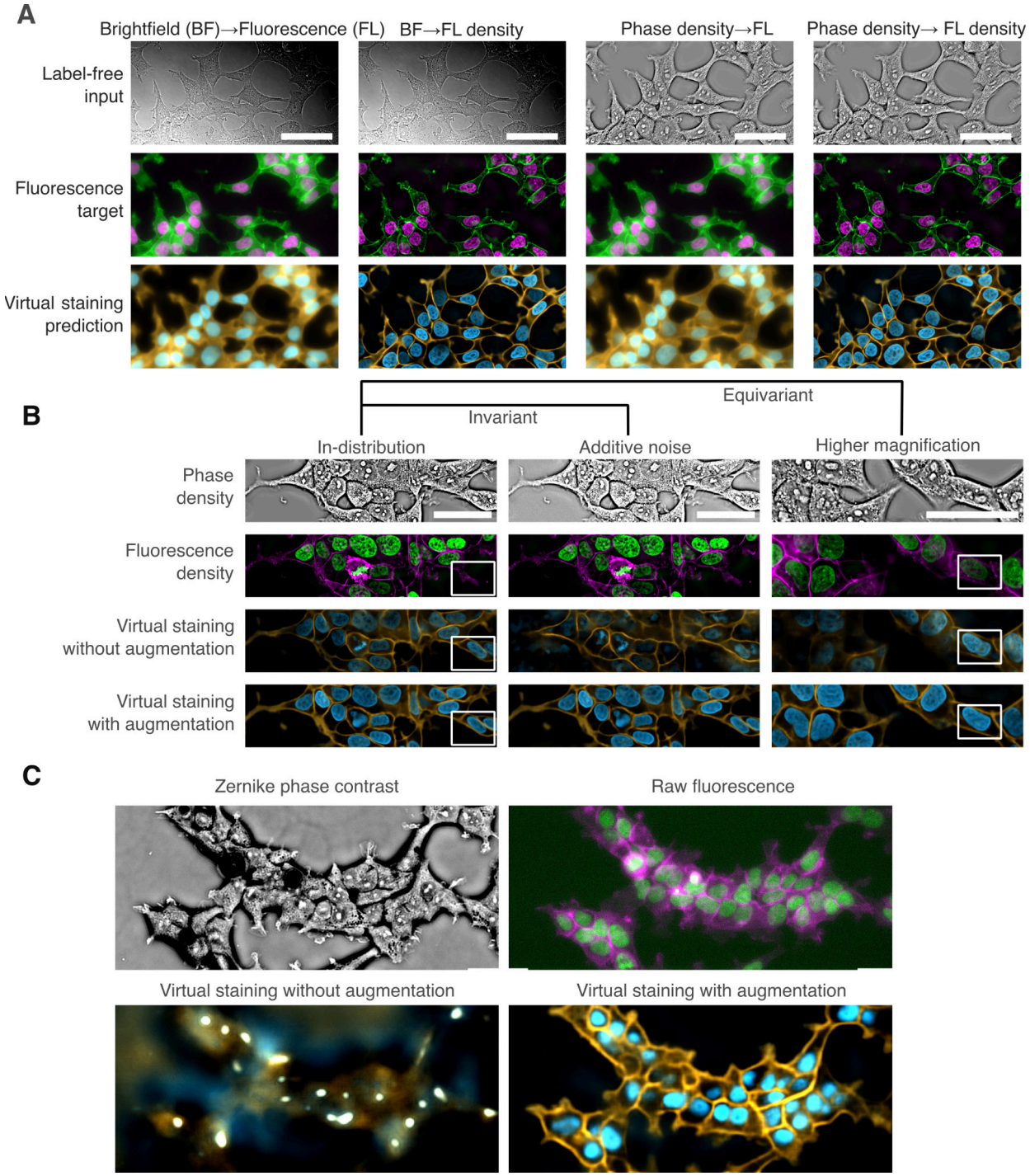
Robust virtual staining:

(A) Schematic illustrating the training (blue arrows) and inference (orange arrows) processes of robust virtual staining models using UNeXt2 with physically informed data augmentations to enhance performance and generalizability. The models virtually stain nuclei and membranes, allowing for single-cell phenotyping without experimental staining. We utilize generalist segmentation models to segment virtually stained nuclei and membranes.

(B) Input phase images (top row), experimental fluorescence images of nuclei and membrane (middle row), and virtually stained nuclei and membrane (bottom row) using VSCyto3D (HEK293T cells), fine-tuned VSNeuromast (zebrafish neuromasts), and VSCyto2D (A549 and BJ-5ta cells). Virtually and experimentally stained nuclei and membranes are segmented using the same Cellpose model. The instance segmentations are compared using the average precision at IoU of 0.5 (AP@0.5). Scale bars: 25 µm.

(C) We rank and refine models based on application-specific metrics, in addition to instance segmentation metrics. (1) Morphological Measurements: We compare cell area in HEK293T cells measured with segmentation of experimentally and virtually stained membranes. (2) Nuclei Count: We compare the number of nuclei in neuromasts identified from experimentally and virtually stained nuclei over short and long developmental time windows. The plot shows the number of nuclei over one hour, measured every 5 minutes on 3 dpf fish.

Figure 2



Deconvolution and data augmentation strategies make the VSCyto3D model robust to label-free imaging parameters:

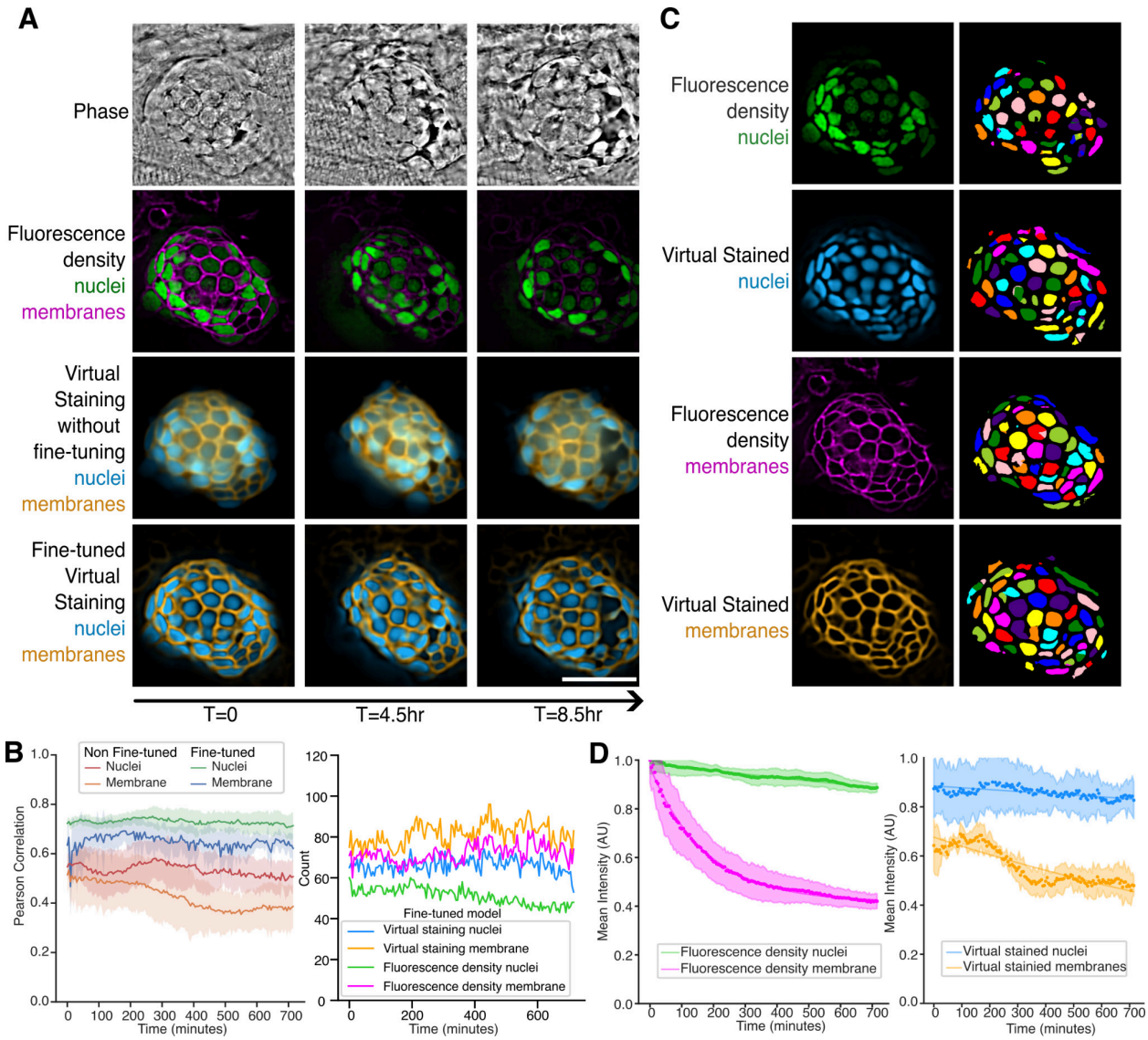
(A) Physics-based reconstructions enhance contrast for virtual staining. Top to bottom: label free input, fluorescence target, and virtual staining prediction. Models are trained

on pairs of raw or reconstructed label free and fluorescence contrast modes. Scale bars: 50 µm.

(B) Predictions of nuclei and membrane from phase image (1st row) using models trained without augmentations (3rd row) are inconsistent with experimental ground truth (2nd row), especially in the presence of noise (center column) or at a different magnification (right column). The predictions using the models trained using spatial and intensity augmentations (see text for details) are invariant to noise and equivariant with magnification. The white box in the in-distribution column highlights the rescue of the lost fluorescence label. The white box in the higher magnification column shows that the model with augmentations correctly predicts one large nucleus whereas the model trained without augmentation predicts two smaller nuclei. Scale bars: 50 µm.

(C) Data augmentation improves generalization to unseen modality. Virtual staining models were trained to predict fluorescence density from phase density and then used to predict nuclei and plasma membrane from Zernike phase contrast image (top left). The correlative raw fluorescence image (top right) shows low signal-to-noise ratio due to light loss in the phase contrast objective. Scale bars: 50 µm.

Figure 3



Generalizing the VSNeuromast model across zebrafish development:

(A) Phase (1st row), experimentally stained nuclei and membrane (2nd row), and virtually stained nuclei and membrane using a model that was not fine-tuned (3rd row) and using a model that was fine-tuned on the subsampled movie (4th row) are shown. We show 3 samples from a 12-hour movie starting at 4 days post-fertilization (4 dpf) and microscope. The VSNeuromast model was fine-tuned with subsampled 4dpf movie (1/12 timepoints). Virtual staining rescues missing nuclei and provides a more accurate read-out of the cell count and their locations than experimental staining. (Scale bar 25um)

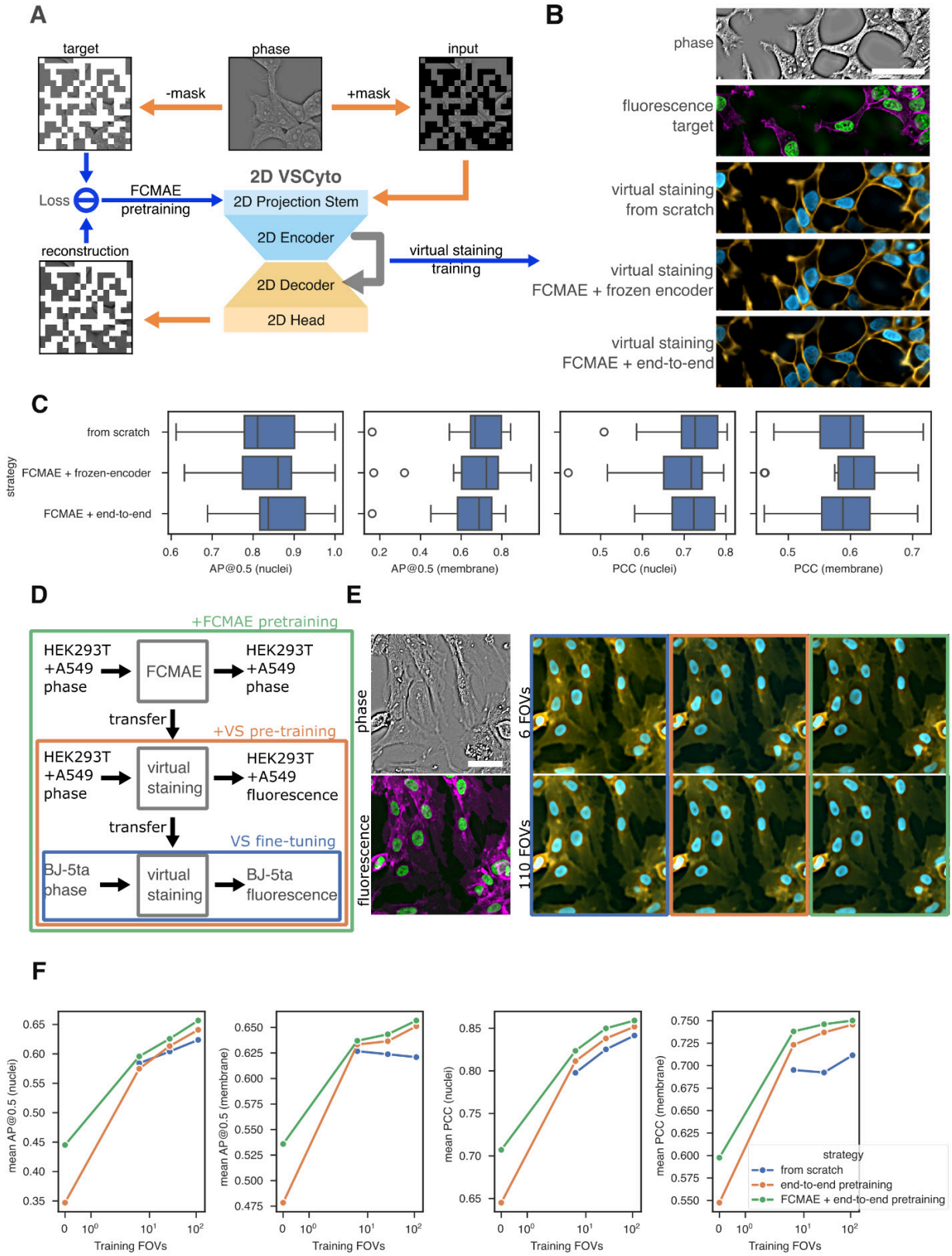
(B) Comparison of nuclear and membrane 3D segmentations predicted using the fine-tuned VSNeuromast Cellpose model. The same segmentation model was applied

consistently to fluorescence density and virtual staining volumes. The VSNeuromast and Cellpose segmentation combination predicts the cell counts with high accuracy. The model's performance was quantitatively assessed using Pearson correlation plots across five neuromasts from the lateral line, comparing both fluorescence density and virtual staining results for nuclei and membranes to highlight the precision of the fine-tuned model.

(C) Segmentation of nuclei and membranes using fine-tuned Cellpose model on experimental fluorescence and virtual staining. Virtual staining reduces over- and under-segmentations, enhancing accuracy compared to experimental fluorescence.

(D) Mean photobleaching curves across five neuromasts showing experimental fluorescence (left) and virtual staining (right) nuclei and membrane pairs. The shaded region indicates the variation in the mean intensity of nuclei in a given frame.

Figure 4


A
target
phase
input
-mask
+mask
Loss
FCMAE
pretraining
2D VSCyto
2D Projection Stem
2D Encoder
2D Decoder
2D Head
reconstruction
virtual staining
training
B
phase
fluorescence
target
virtual staining
from scratch
virtual staining
FCMAE + frozen encoder
virtual staining
FCMAE + end-to-end
C
strategy
from scratch
FCMAE + frozen-encoder
FCMAE + end-to-end
AP@0.5 (nuclei)
AP@0.5 (membrane)
PCC (nuclei)
PCC (membrane)
D
+FCMAE pretraining
HEK293T
+A549
phase
FCMAE
HEK293T
+A549
phase
transfer
+VS pre-training
HEK293T
+A549
phase
virtual
staining
HEK293T
+A549
fluorescence
transfer
VS fine-tuning
BJ-5ta
phase
virtual
staining
BJ-5ta
fluorescence
E
phase
fluorescence
6 FOVs
110 FOVs
F
mean AP@0.5 (nuclei)
mean AP@0.5 (membrane)
mean PCC (nuclei)
mean PCC (membrane)
Training FOVs
strategy
from scratch
end-to-end pretraining
FCMAE + end-to-end pretraining

Few-shot generalization of the VSCyto2D model to new cell types:

(A) The encoder of UNeXt2 is pre-trained with a fully convolutional masked autoencoder (FCMAE) recipe to enable generalized feature encoding without paired data. The model can then be fine-tuned for virtual staining on a specific cell type.

(B) Virtual staining of nuclei and membrane in HEK293T using models trained from scratch or pre-trained (FCMAE) and then fine-tuned on HEK293T data. pre-training improves the high-frequency features in predictions. Scale bar: 50 μm. PCC: Pearson correlation coefficient. AP@0.5: average precision at IoU =0.5.

(C) The pre-training protocol has similar segmentation and regression performance for a single cell type.

(D) Flow chart of 3 training strategies to generalize the virtual staining model pre-trained with A549 and HEK293T cells to the BJ-5ta cell type with limited training samples. The bounding boxes indicate strategies: (blue) virtual staining pre-training from scratch with paired images of BJ-5ta; (orange) pre-training with paired images of HEK293T and A549, and fine-tuning with paired images of BJ-5ta; and (green) FCMAE pre-training with only the phase images of HEK293T and A549, virtual staining pre-training with images of HEK293T and A549, and fine-tuning with paired images of BJ-5ta. The pre-training steps initialize model weights in the encoder (FCMAE) and decoder (virtual staining) of UNeXt2.

(E) Virtual staining images of nuclei and membrane in BJ-5ta using 3 models described in D. Performance scales with the increasing number of BJ-5ta FOVs used for fine-tuning. Scale bar: 50 μm.

(F) AP@0.5 of segmented nuclei and membrane as a function of the number of fields of view used for the test dataset used for training strategies shown in D. The pre-trained models show superior performance scaling relative to the number of BJ-5ta FOVs used for fine-tuning. Pre-trained models fine-tuned with less data can match or outperform models trained with more data from scratch.

Tables

Table 1

Model-Z(in, out)	Prediction target	Training XY patch size (pixel)	Training epochs	Multiply-Adds (G)	Params (M)	PCC (nuclei)	AP@0.5 (nuclei)
2.5D UNet-Z(5, 1)	Nuclei	512	100	1005.51	2.01	0.721	0.804
2.5D UNet-Z(5, 1)	Nuclei and plasma membrane	512	100	1006.12	2.01	0.720	0.829
2.5D UNet-Z(5, 1)	Nuclei and plasma membrane	384	50	1006.12	2.01	0.720	0.733
UNeXt2-Z(5, 5)	Nuclei and plasma membrane	384	50	723.62	32.04	0.714	0.854

Comparison of the computational complexity, capacity, and performance of UNeXt2 models with previously published 2.5D UNet model: We compare metrics of accuracy of regression (Pearson correlation coefficient, PCC) and instance segmentation (average precision at IoU threshold of 0.5, AP@0.5) of nuclei on the central slice of HEK293T images. The computational complexity is measured with the number of multiply-add operations during inference on ZYX input size (5, 2048, 2048) with batch size 1. The model capacity is measured with the number of learnable parameters. Predicting both nuclei and membrane targets improves nuclei prediction with 2.5D UNet. UNeXt2 architecture provides higher learning capacity than 2.5D UNet architecture at similar computational complexity. Predictions with UNeXt2-Z(5,5) are shown in Figure 1.

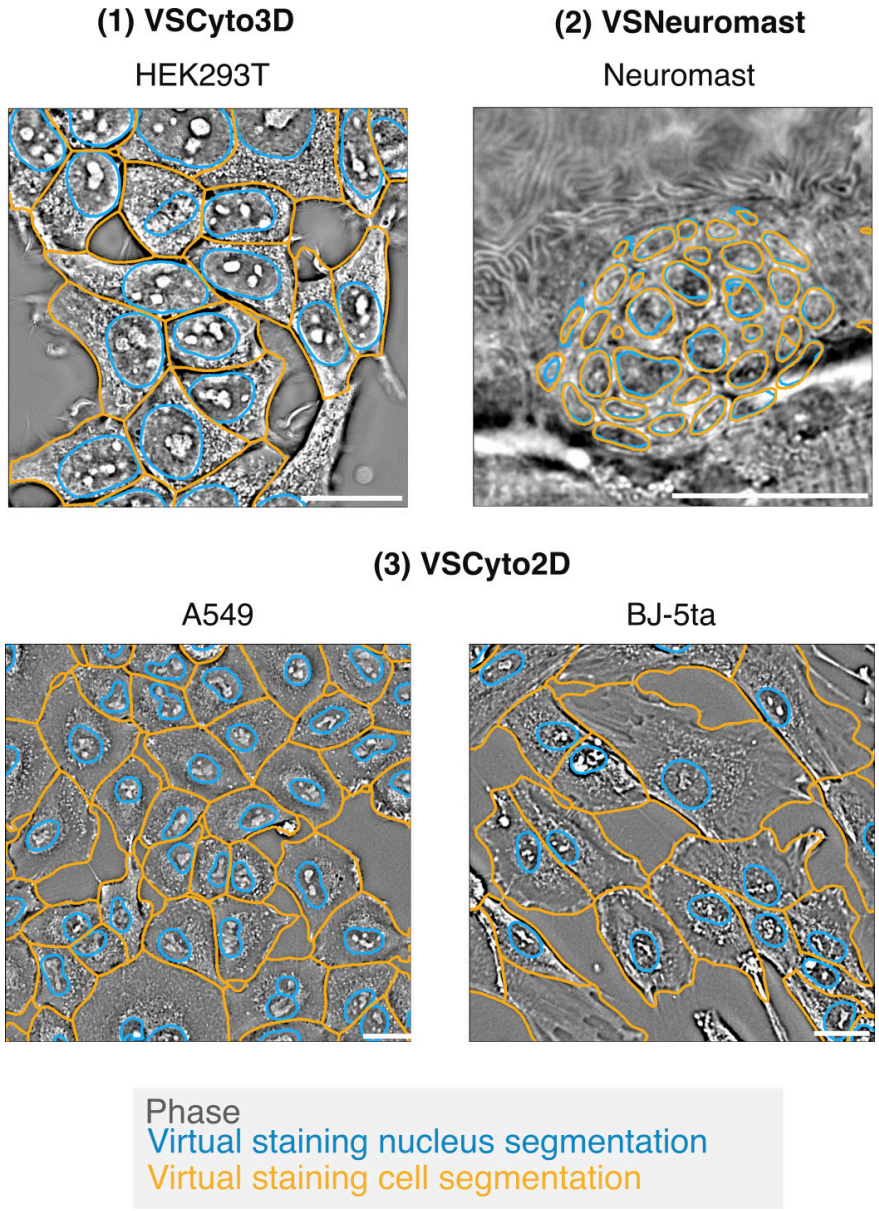
Table 2

Model-Z(in, out)	Loss	MSE	2D-SSIM	mIOU	AP	AR
2.5D UNet-Z(5, 1)	MSE	4.16	0.373	0.752	0.441	0.629
2.5D UNet-Z(5, 1)	Mixed	3.86	0.513	0.794	0.533	**0.655**
UNeXt2-Z(5, 5)	MSE	4.00	0.432	0.775	0.494	0.613
UNeXt2-Z(5, 5)	Mixed	**3.77**	**0.620**	**0.799**	**0.537**	0.645

Comparison of the performance of models trained with 2 loss functions on the center slice of the HEK293T test dataset. The models trained with the mixed loss also have lower test MSE than models trained with the MSE loss.

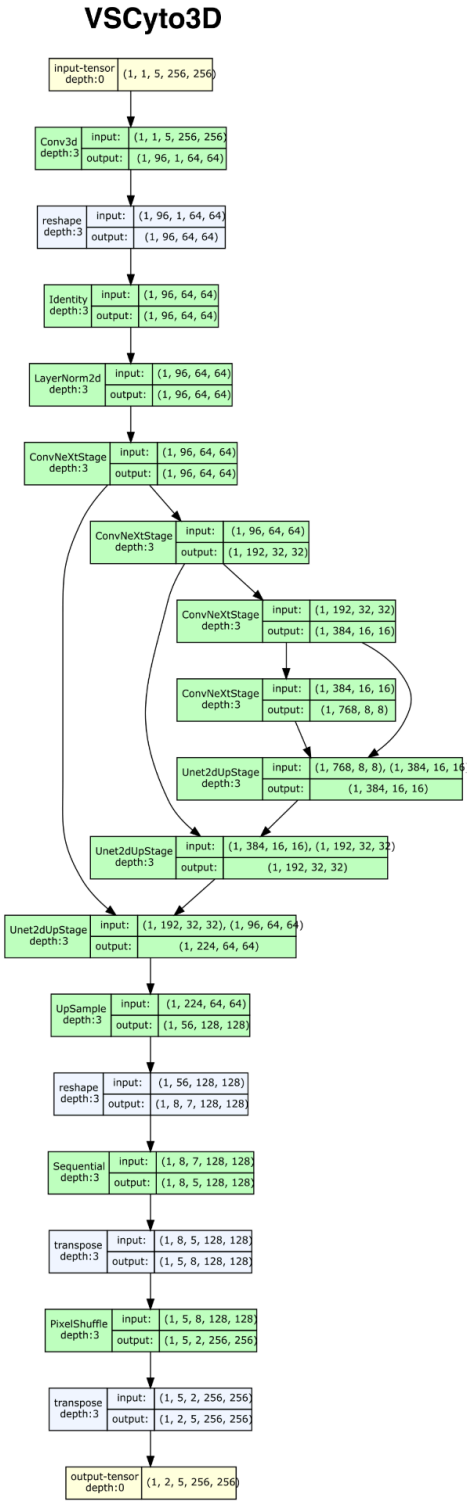
Supplementary Materials

Figure S1



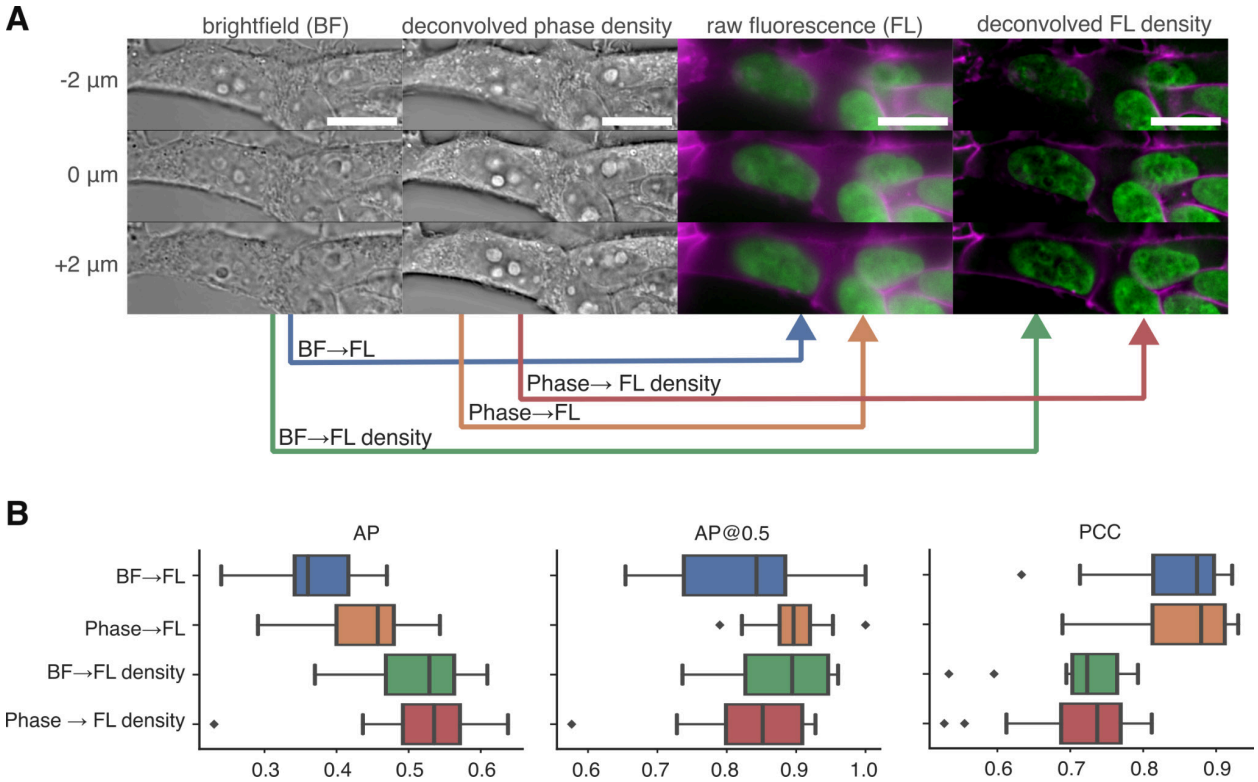
Examples of nuclei and cell segmentation from virtual stained images overlaid on the corresponding phase images.

Figure S2



The model architecture used to train VSCyto3D

Figure S3

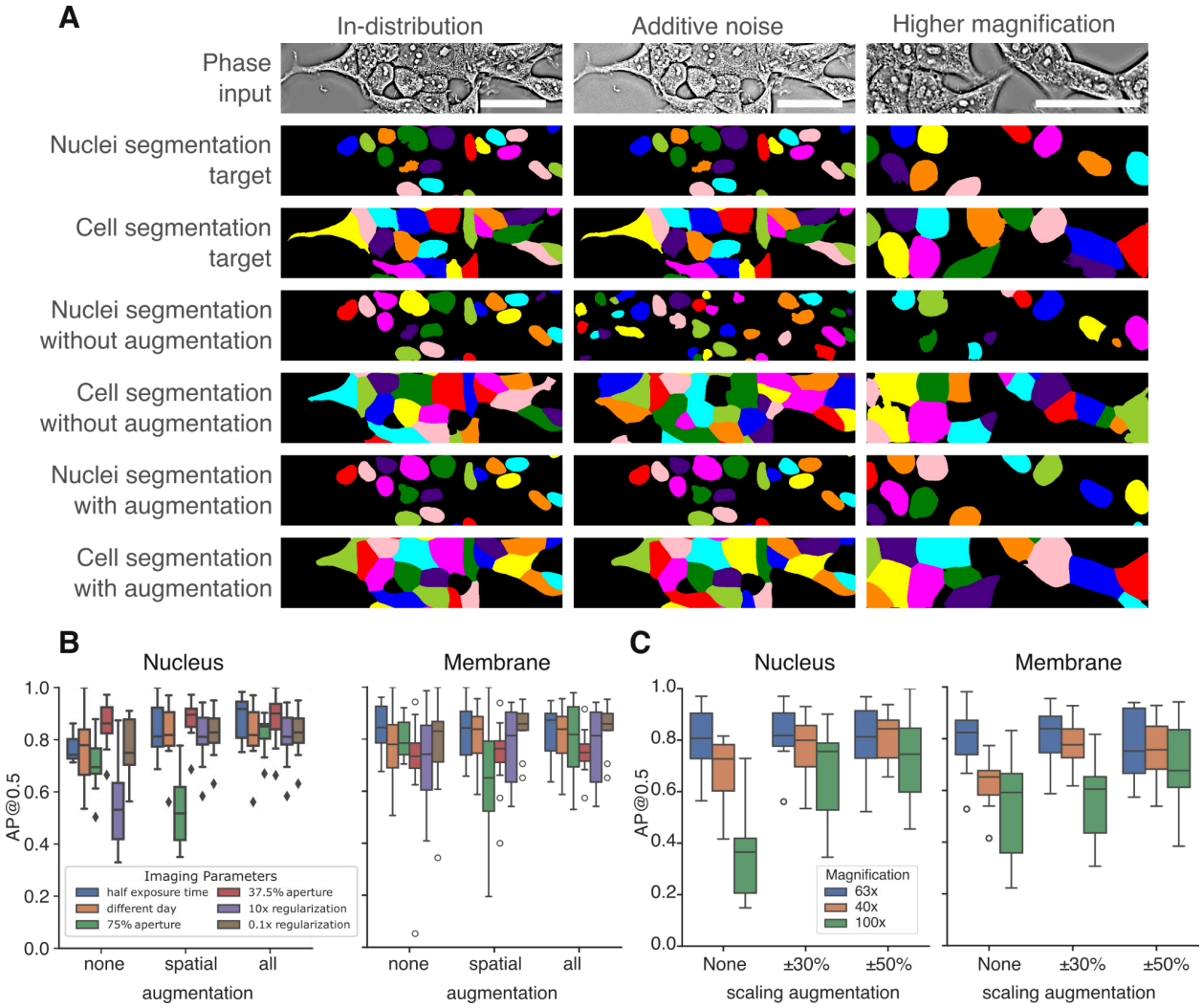


Deconvolution improves contrast of fine features for virtual staining:

(A) Comparison of contrast in brightfield (BF) and fluorescence (FL) images with the corresponding deconvolution of the phase density and fluorescence density. We trained four virtual staining models that translate between raw and deconvolved versions of label-free and fluorescence contrasts (indicated by arrows). Scale bars: 10 µm.

(B) The average precision (AP) and average precision at IoU =0.5 (AP@0.5) for nuclei segmented from experimentally and virtual stained images are shown. Virtually stained images were predicted with four models indicated on the y-axis. Instance segmentations of experimentally stained nuclei were proofread manually. Deconvolution of BF and FL volumes leads to more accurate segmentation of nuclei. We also assess how the phase and fluorescence density, deconvolved from brightfield (BF) and fluorescence (FL) volumes, respectively, affect the Pearson cross-correlation (PCC).

Figure S4



Augmentation-induced robustness in virtual staining improves segmentation of nuclei and cells:

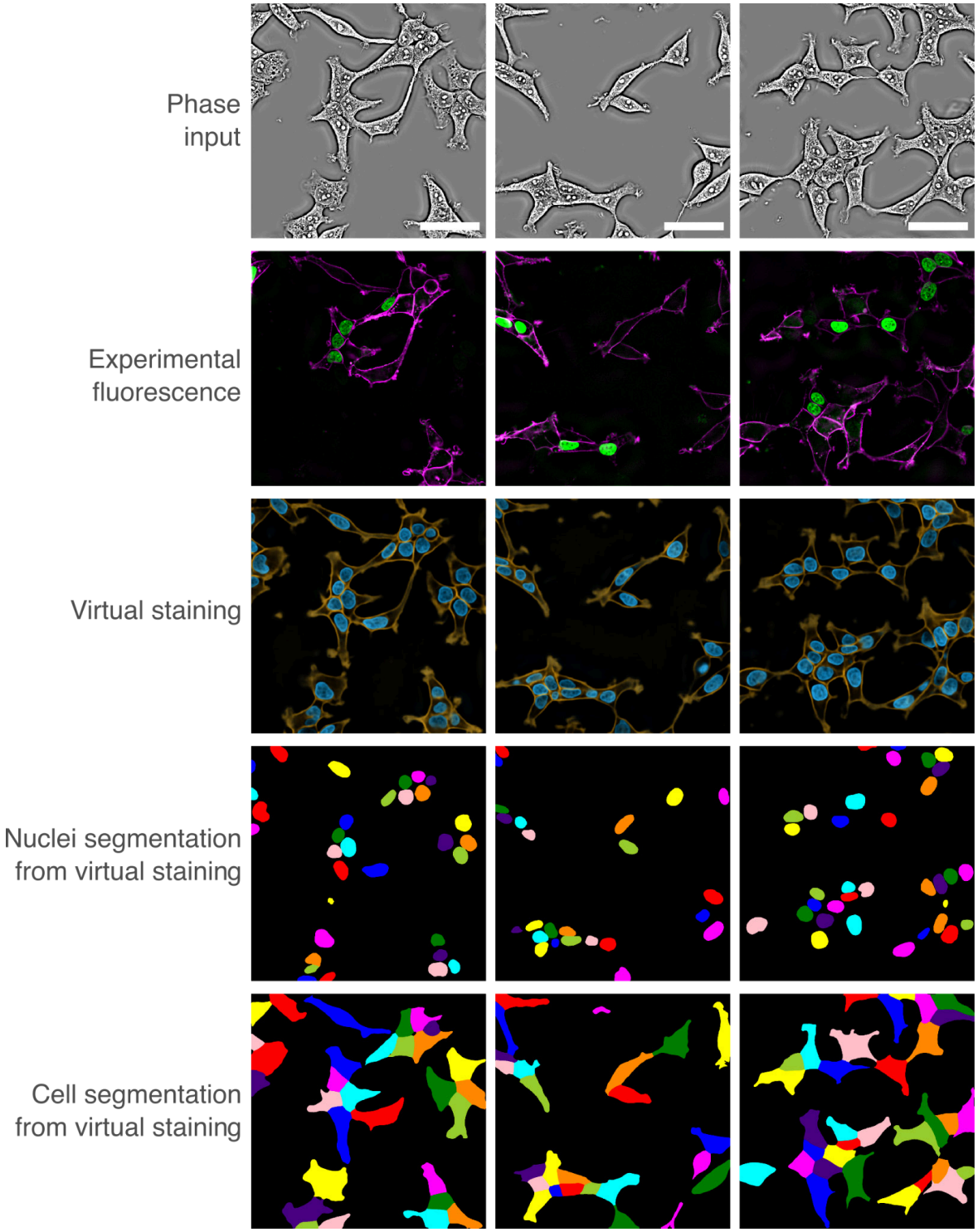
(A) Segmentations from virtual staining predictions of VSCyto3D models trained without augmentation are inaccurate when noise is added to the phase image or when phase image at a different magnification is used for inference. The segmentations are more reliable when using augmentations as described in Methods. Scale bars: 50 µm.

(B) Test FOVs were acquired with diverse imaging parameters (see the legend). Virtual staining models trained without augmentation (group: none), with spatial augmentation (group: spatial), and with spatial and intensity augmentations (group: all) were used to virtually stain nuclei and membranes in test FOVs. We compare the AP@0.5 between instance segmentations from virtually stained nuclei and proofread instance segmentations from experimentally acquired nuclei images. The solid lines in the box plot indicate the median and interquartile ranges of AP@0.5 across all FOVs in the test set. The predictions of models become invariant to changes in the imaging parameters

as more data augmentations are used during training. AP@0.5 for membrane segmentation also indicates that the augmentations make the membrane predictions invariant to imaging parameters.

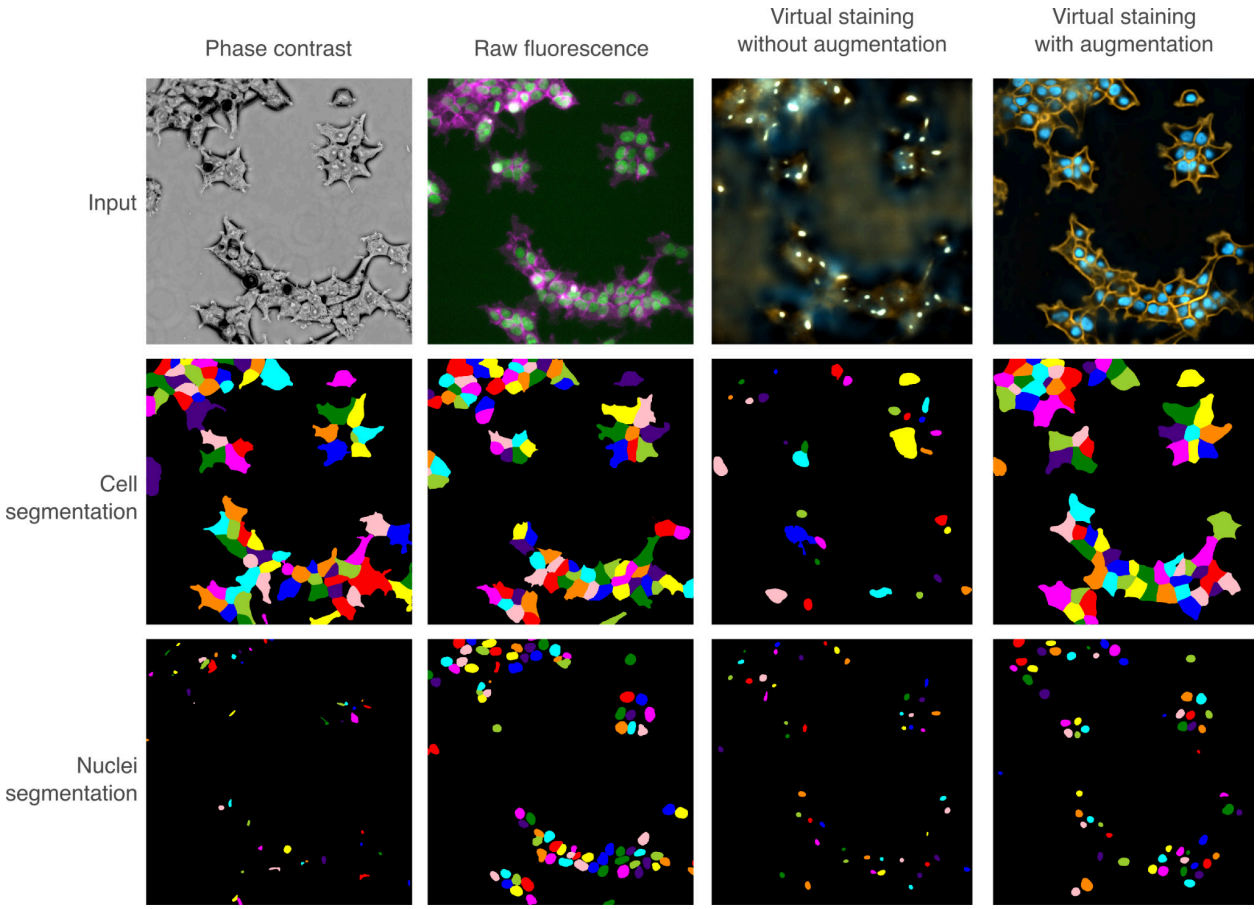
(C) AP@0.5 for test FOVs acquired at the same (63x), higher (100x), and lower (40x) magnifications relative to the training dataset, when no scaling augmentation (group: none), and increasing scaling augmentation (groups: ±30%, ±50%) is used when training models for joint virtual staining of nuclei and membrane. The metrics for both nuclei and membranes indicate that the scaling augmentations make the predictions equivariant to magnification.

Figure S5



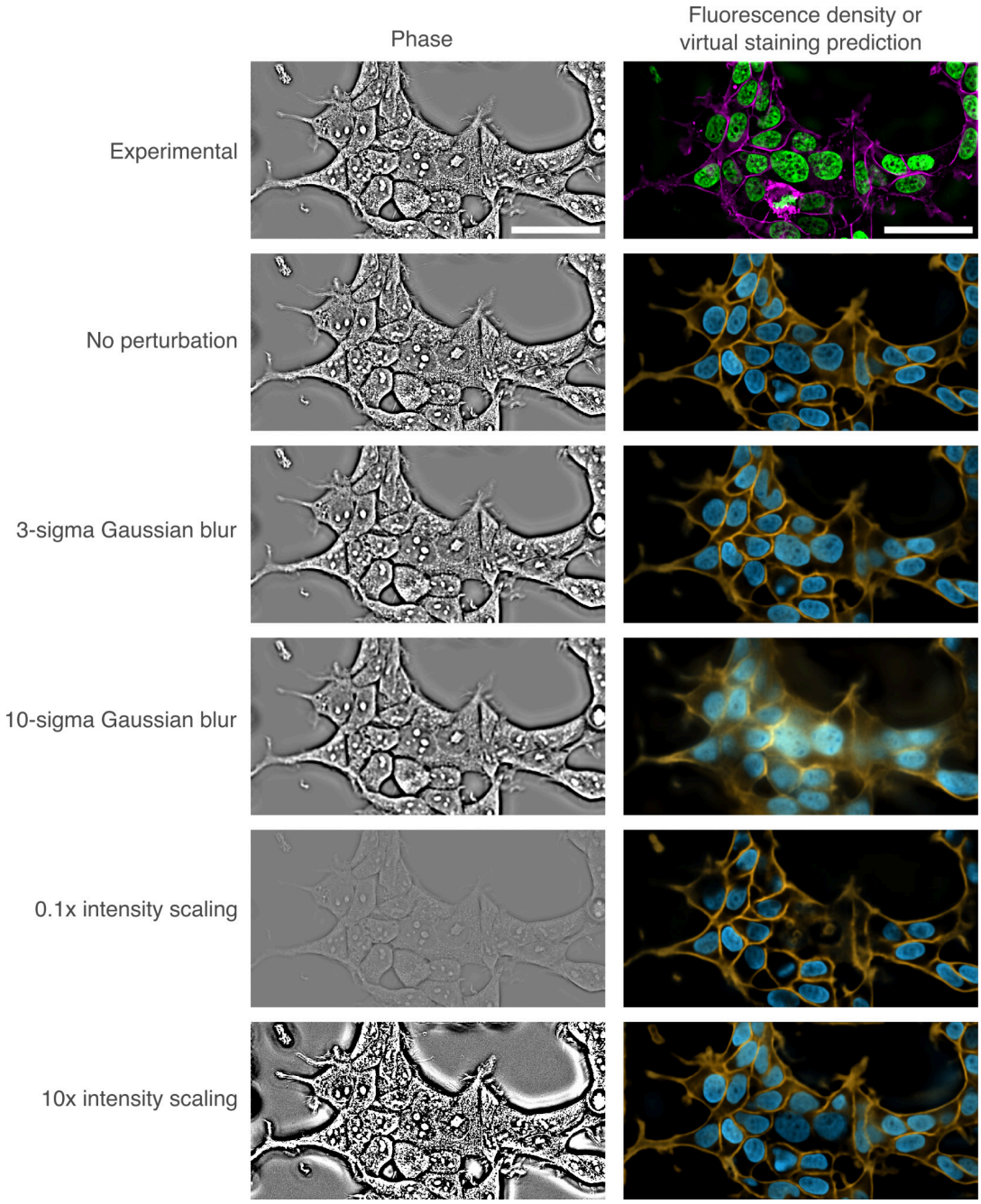
Virtual staining from phase rescues missing fluorescence labels: Experimentally and virtually stained in HEK293T cells nuclei and membrane for 75% aperture and corresponding segmentations. Scale bars: 50 µm.

Figure S6



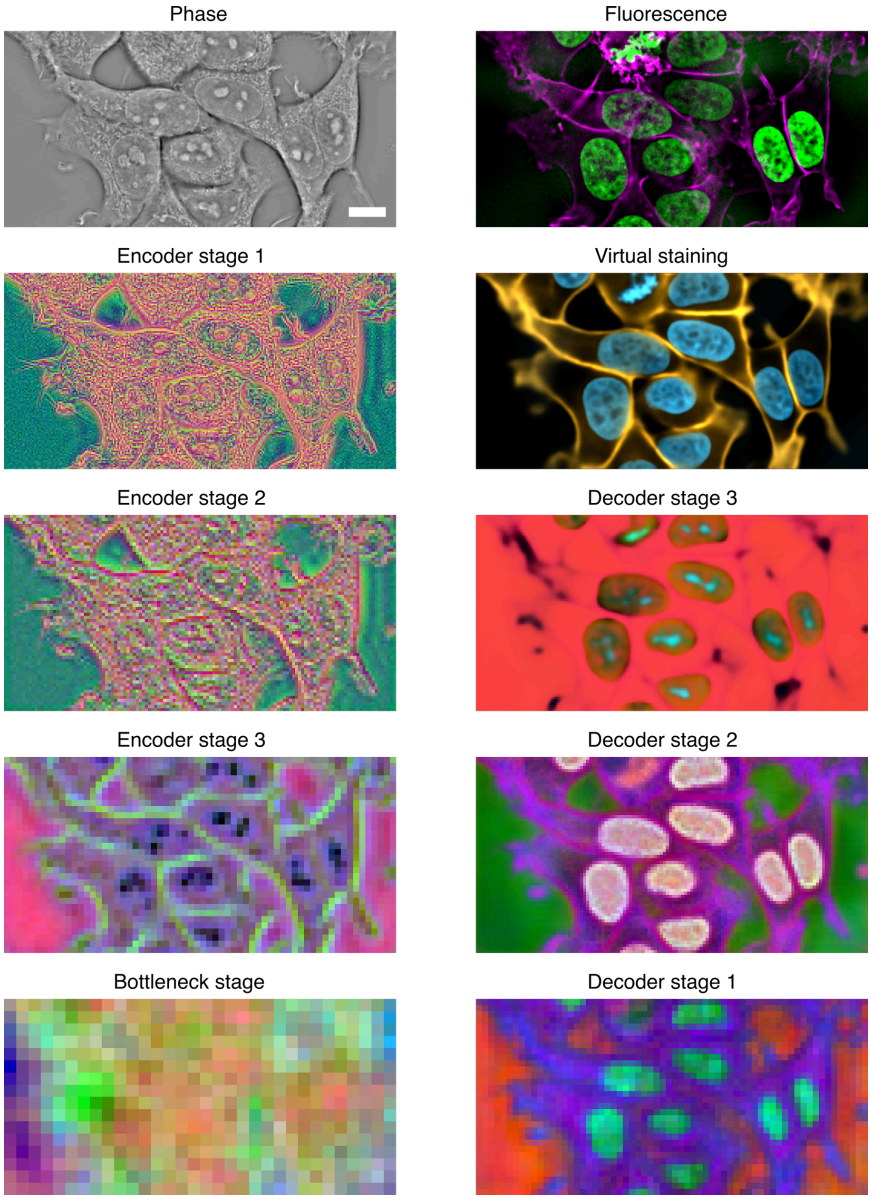
Cell and nuclei segmentation from Zernike phase contrast, raw fluorescence, and virtual staining images. Augmentation improves the virtual staining prediction for segmentation.

Figure S7



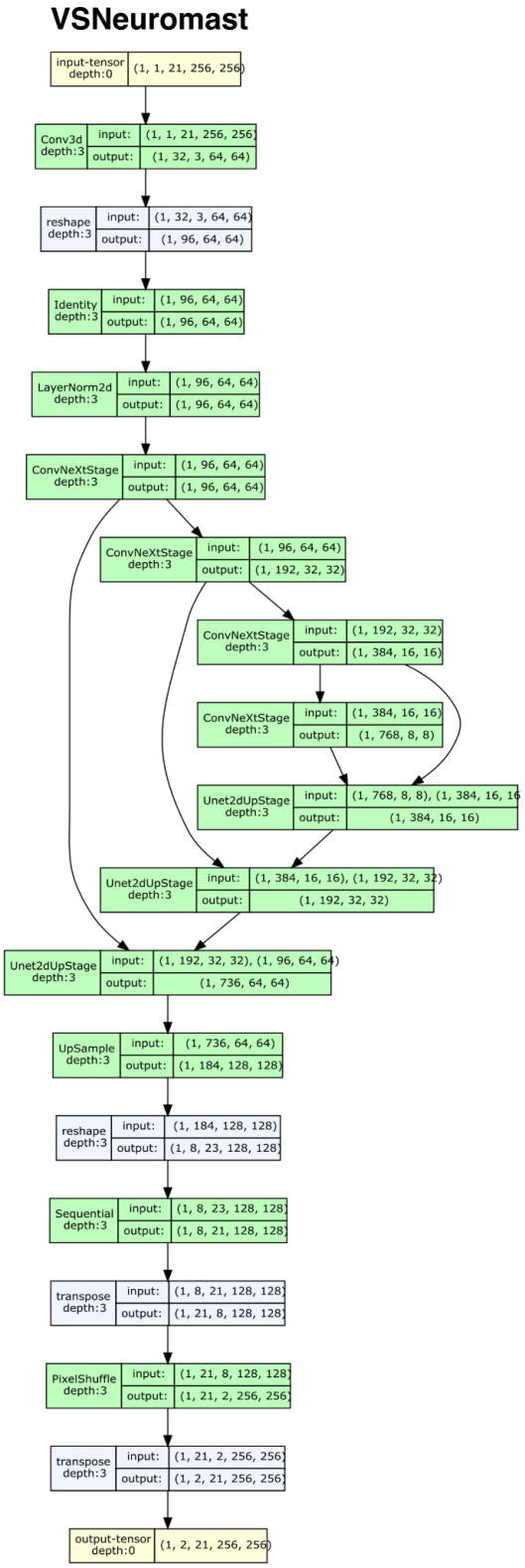
VSCyto3D model's robustness to perturbations in input: Simulated image perturbations are applied to the input phase image. The virtual staining prediction is robust to a certain range of input perturbation and fails when the deviation is too large (σ=10 pixels). For example, when a very strong Gaussian blur is applied to the input, the model cannot differentiate the in-focus slice from a defocus slice of the imaging volume, and predicts a blurry image. Scale bars: 50 μm.

Figure S8



Visualization of features learned by VSCyto3D: Input, prediction, and intermediate feature maps of the 3DVSCyto (Figure 2B 'deconvolved -> deconvolved') model trained on HEK293T cells. The first 3 principal components of the feature map from each ConvNext stage are rendered as RGB values for an illustrative input image patch. Scale bar: 10 µm.

Figure S9



The model architecture used to train VSNeuromast

Figure S10

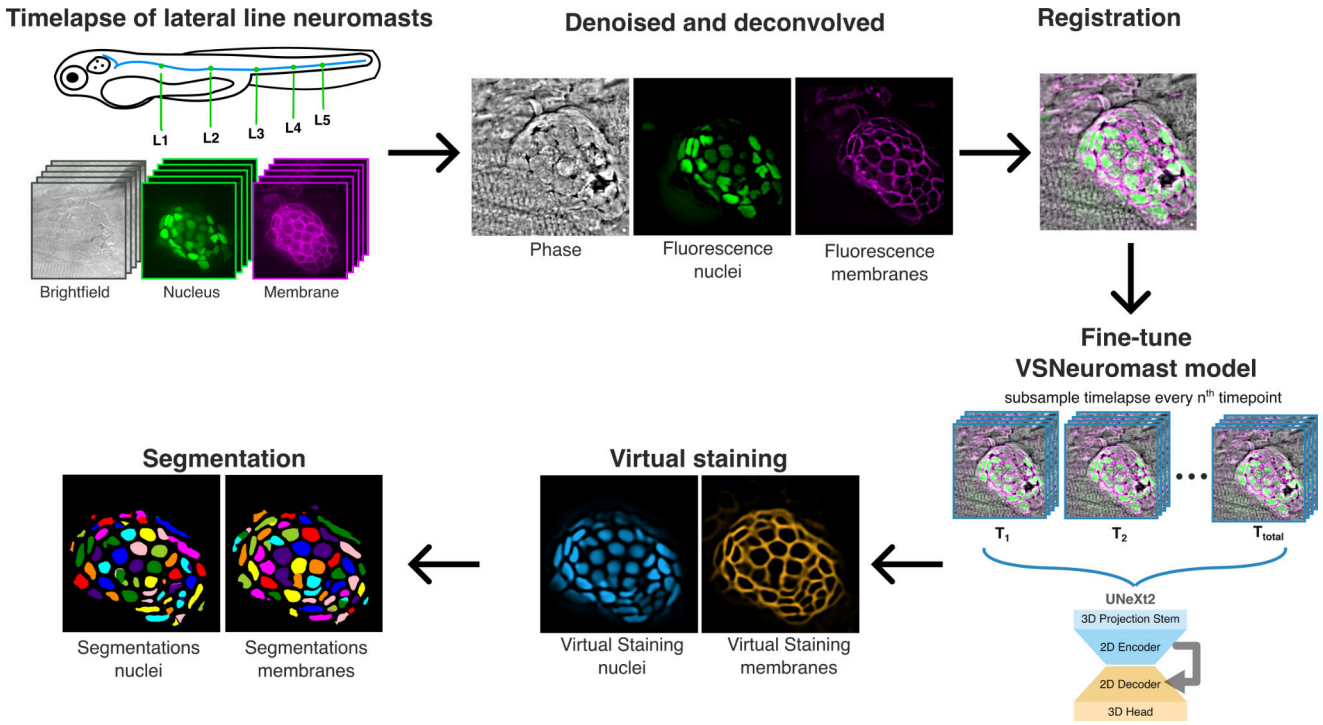
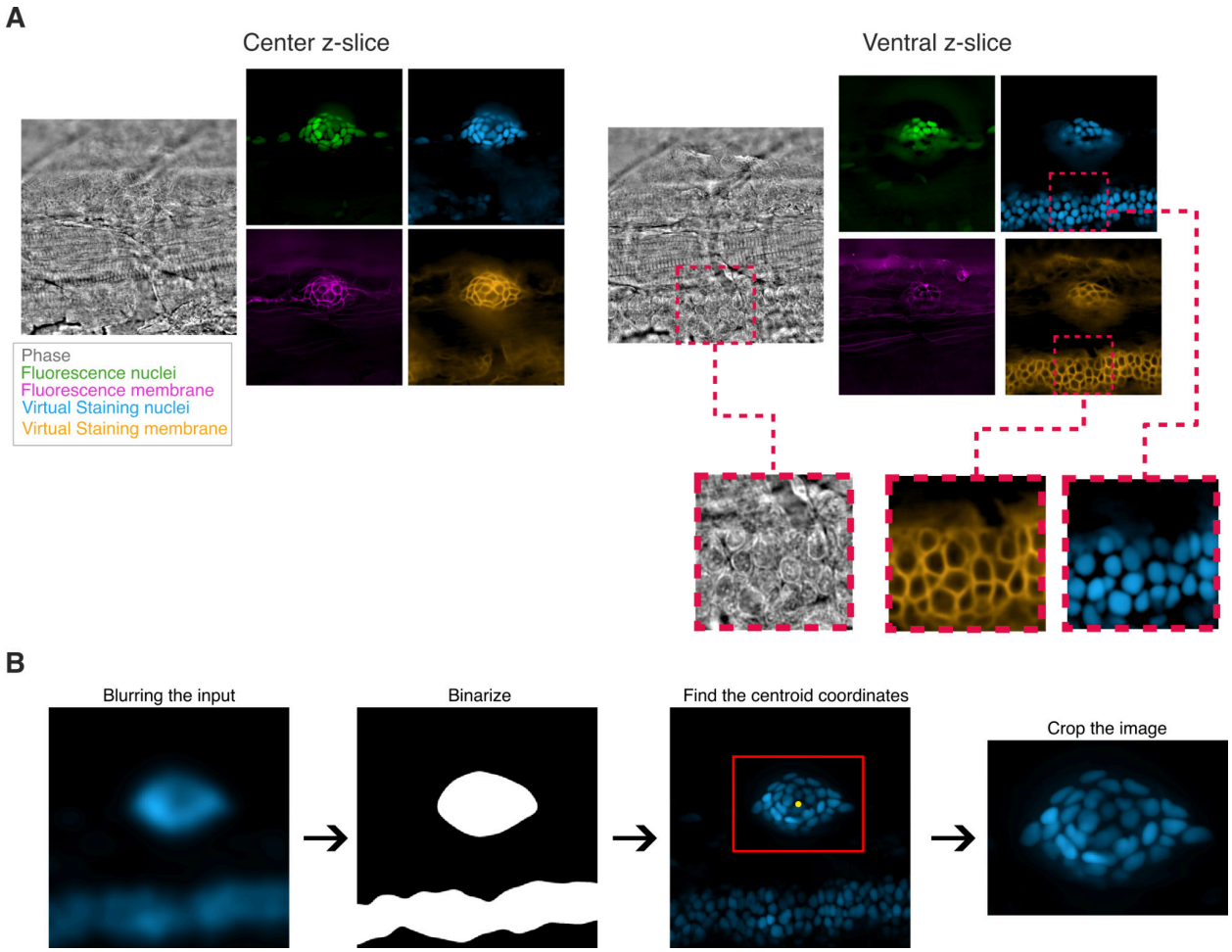


Illustration of the workflow for fine-tuning the VSNeuromast model: The model is fine-tuned to generate accurate predictions and segmentations by subsampling the timelapse and using these time points as training data. Images of neuromasts during preprocessing (denoising, deconvolution, registration) and postprocessing (segmentation) steps are also shown.

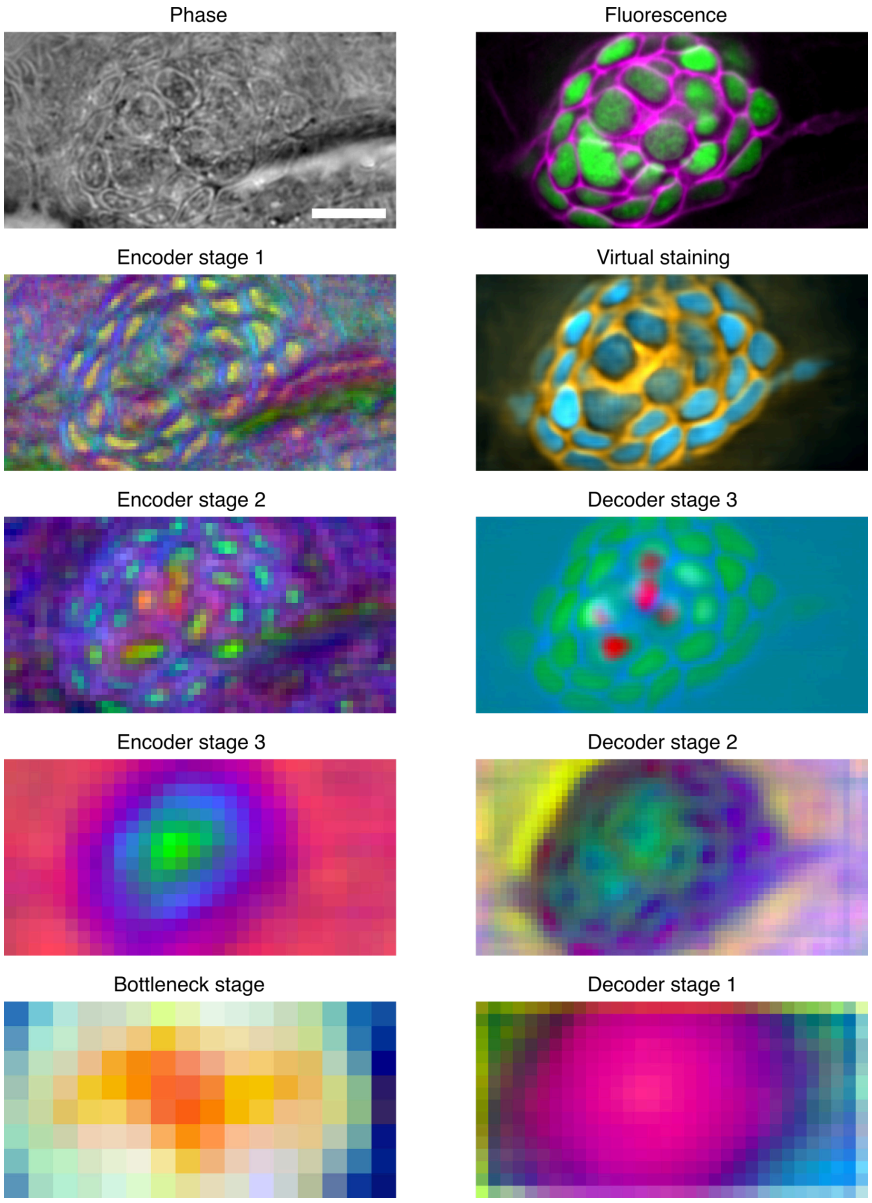
Figure S11



Post-processing is needed to distinguish virtually stained neuromast cells from non-neuromast cells:

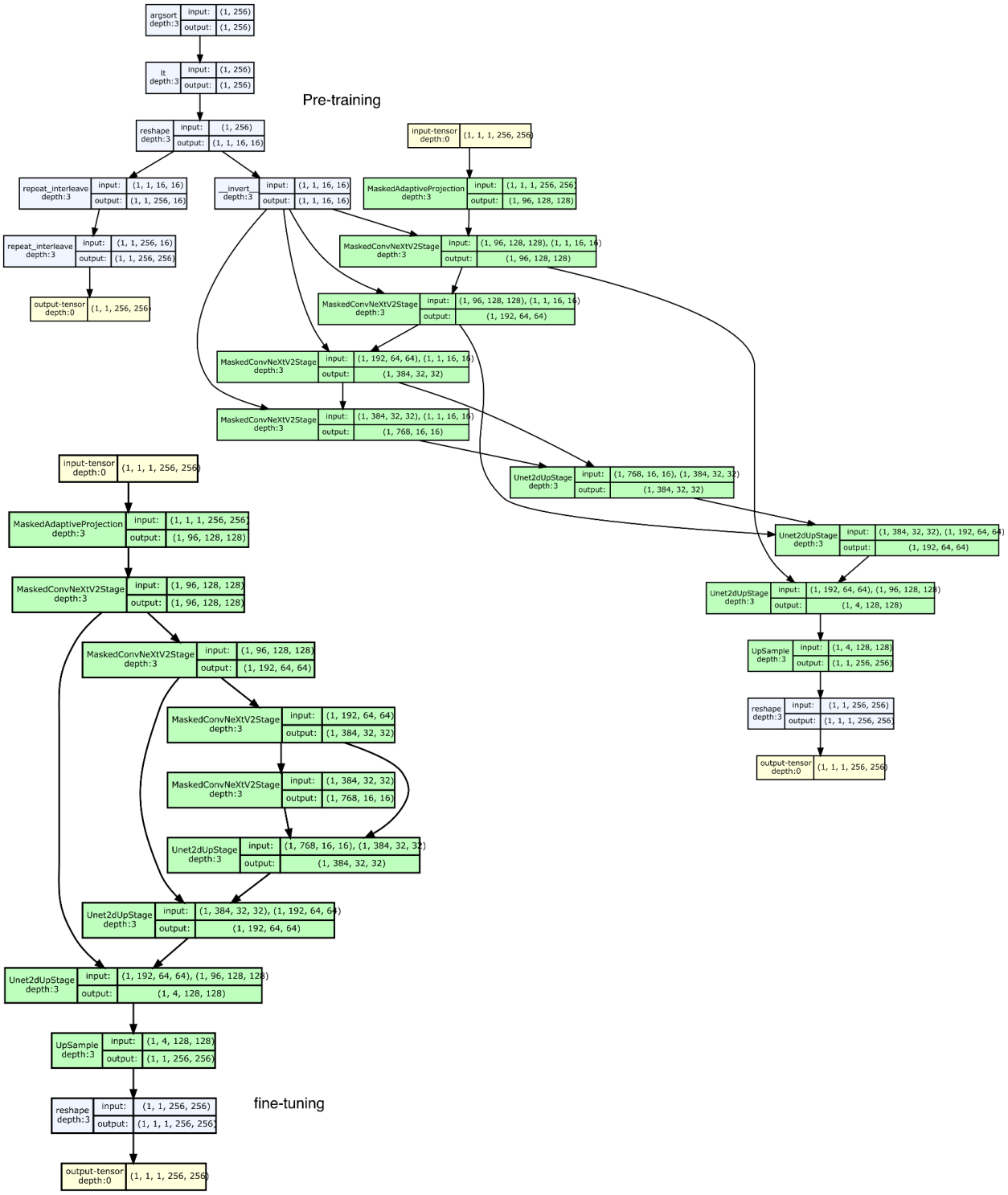
A. Phase, fluorescence and virtual staining pairs of the central and ventral slices depicting how the model generalizes to other cell types with similar morphology.
B. Processing pipeline to isolate the neuromasts from the whole FOV. The pipeline is used for generating the instance segmentations and performance metrics.

Figure S12



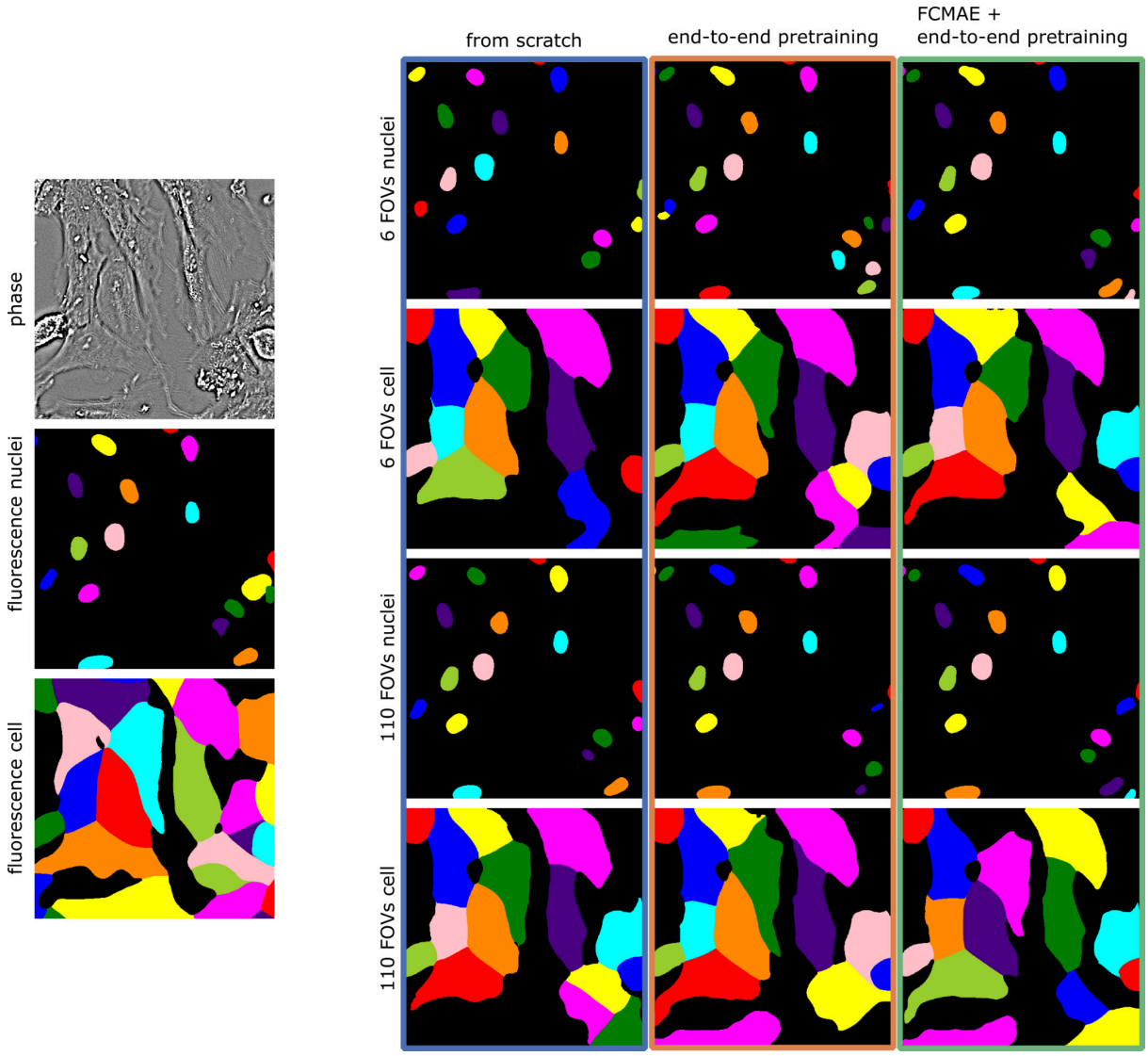
Visualization of features learned by VSNeuromast: Input, prediction, and intermediate feature maps of the 3DVSNeuromast model trained on zebrafish neuromasts. The first 3 principal components of the feature map from each ConvNext stage are rendered as RGB values for an illustrative input image patch. Scale bar: 10 µm.

Figure S13



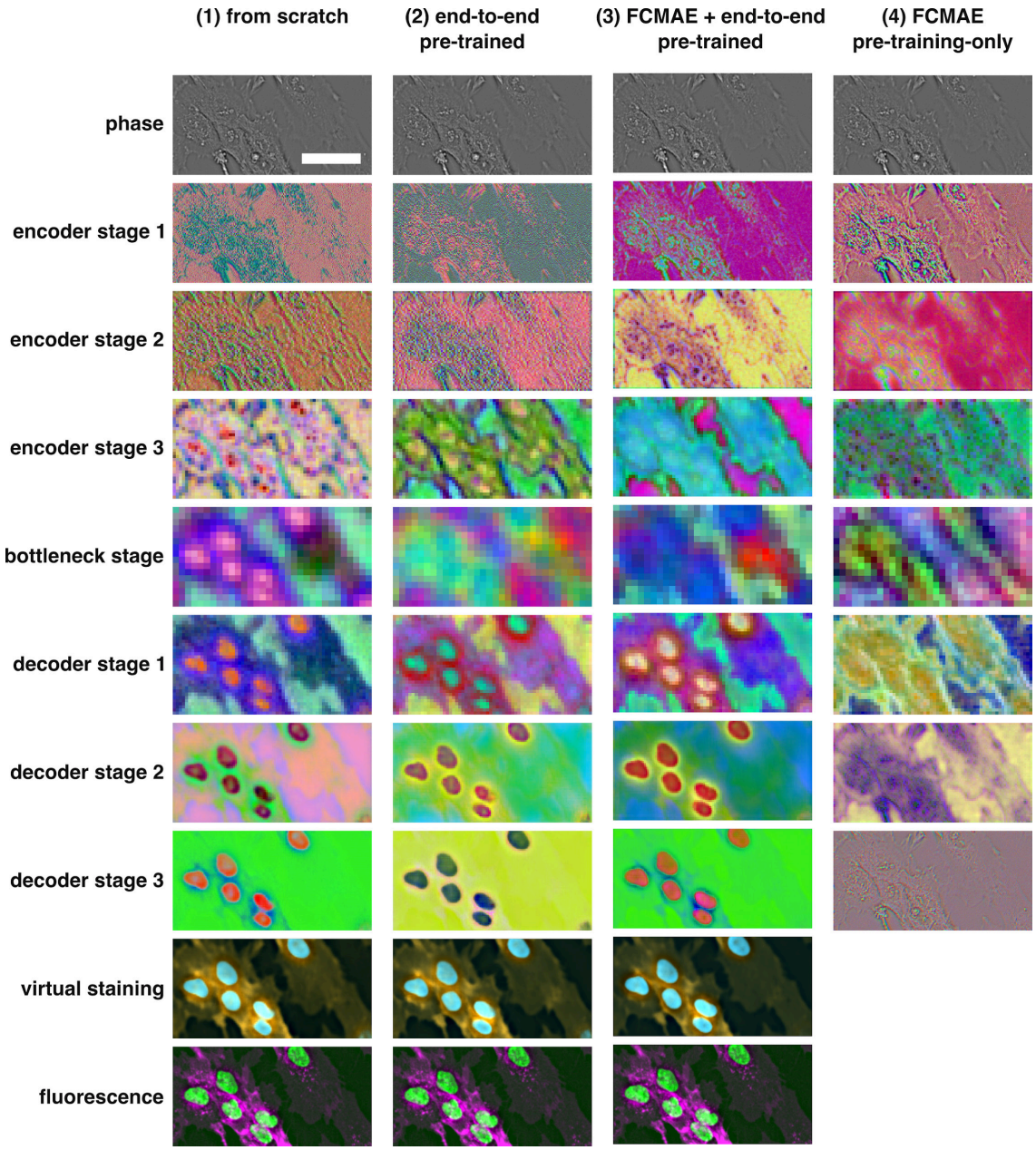
Model architectures used for training VSCyto2D

Figure S14



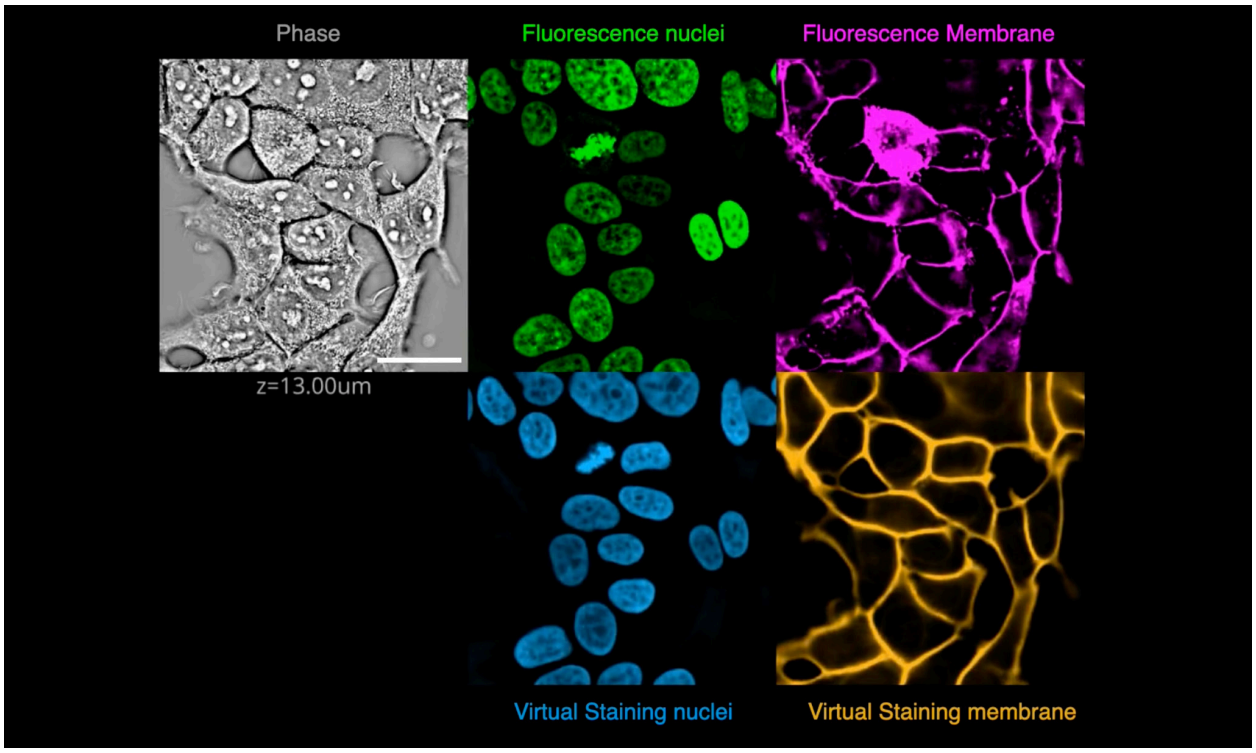
Nuclei and cell segmentation from different virtual staining models. The segmentations are shown for 6 FOVs and 110 FOVs models per training strategy in Figure 4E.

Figure S15



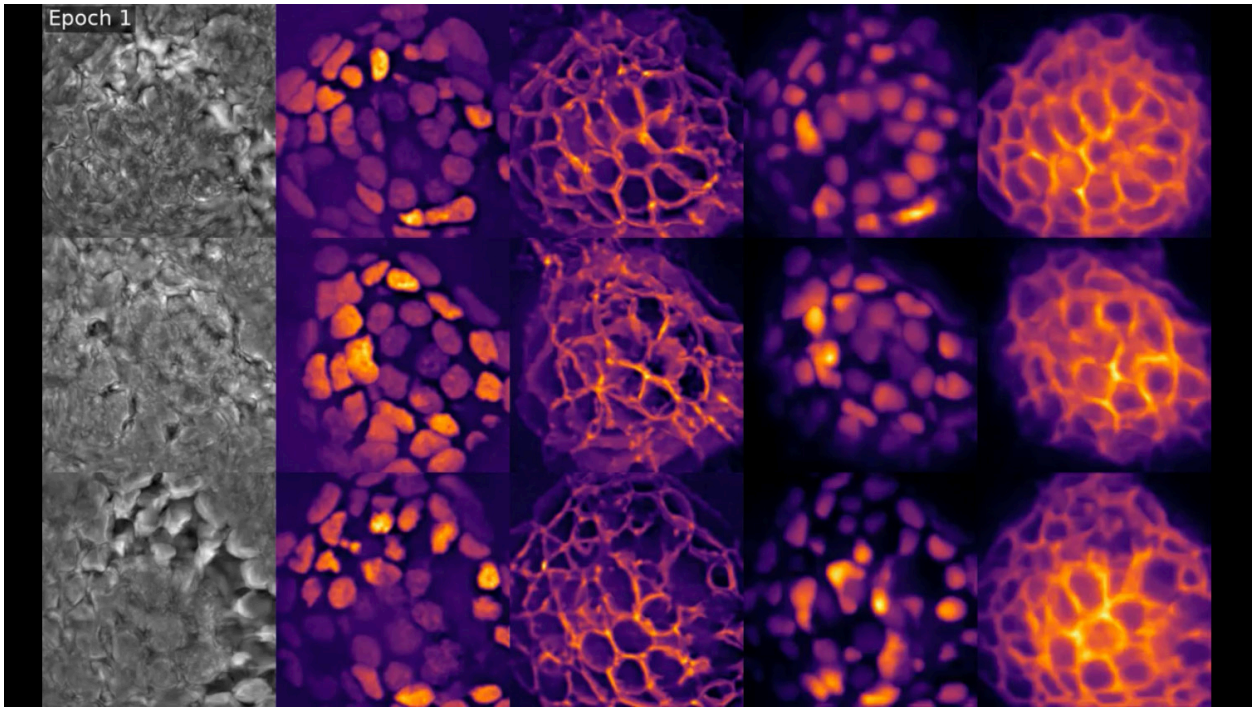
Visualization of features learned by VSCyto2D: Input, prediction, and intermediate feature maps of the 2DVSCyto and FCMAE models. The first 3 principal components of the feature map from each ConvNext stage are rendered as RGB values for an illustrative input image patch. (1) model trained from scratch on BJ-5ta; (2) model pre-trained on virtual staining of HEK-293T and A549, and then fine-tuned on BJ-5ta; (3) model pre-trained with FCMAE and virtual staining of HEK-293T and A549, and then fine-tuned on BJ-5ta; (4) FCMAE model of HEK-293T and A549, not trained for virtual staining. Scale bar: 50 μm.

Video 1



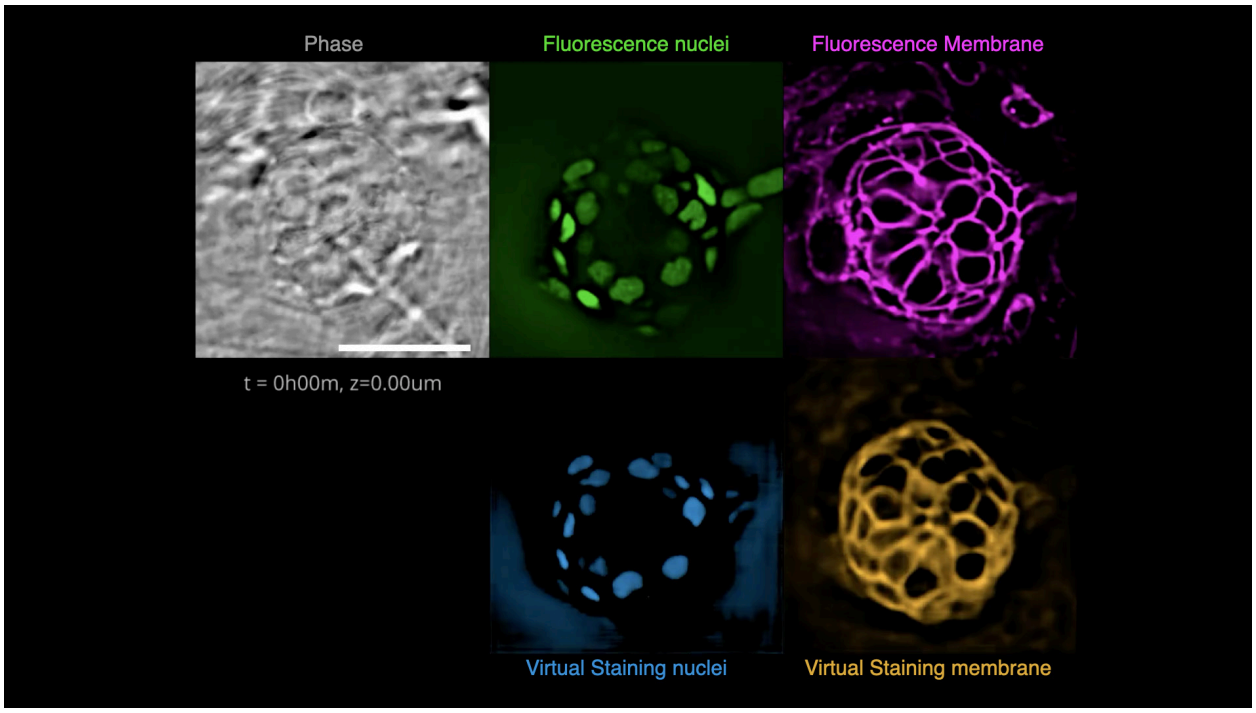
Through-focus movie of HEK293T cells: phase, experimentally stained nuclei (green) and membrane (magenta), virtually stained nuclei and membrane with VSCyto3D. (scale bar 25µm)

Video 2



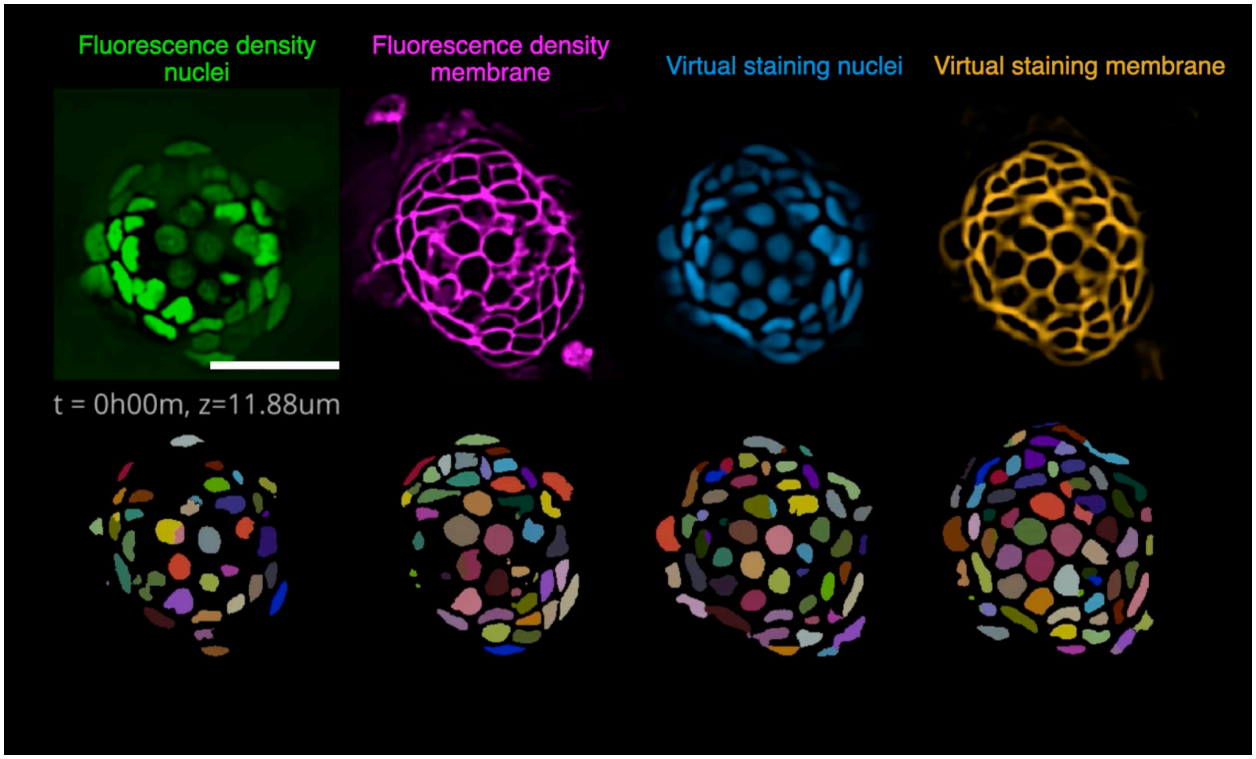
Evolution of neuromast predictions during fine-tuning. The first three columns depict the 2D source (phase) and target (nuclei and membrane) pairs of three different fields of view (FOVs) from the validation dataset. The last two columns feature the virtual staining predictions of nuclei and membrane respectively.

Video 3



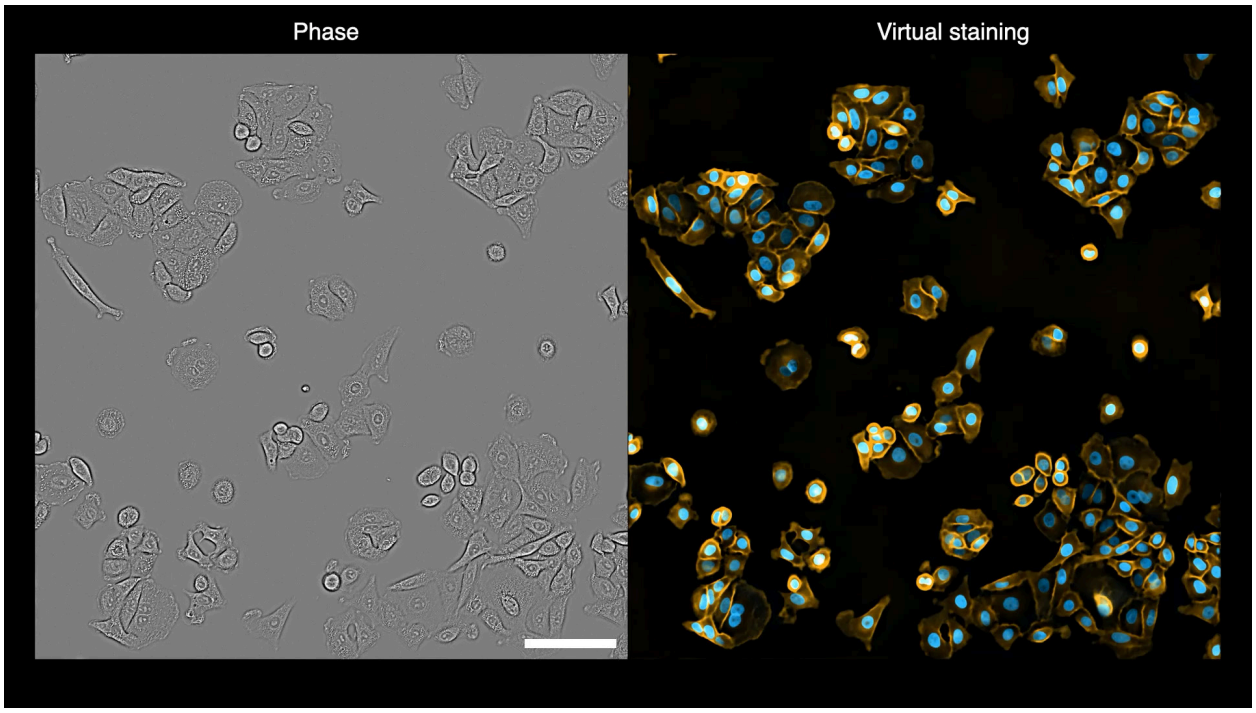
Axial and temporal fly-through of fluorescence and virtually stained landmarks. Displayed are the phase image in gray, fluorescence nuclei in green, and membrane in magenta, along with virtually stained nuclei in blue and membrane in orange, predicted using the fine-tuned VSNeuromast. (Scale bar: 25µm)

Video 4



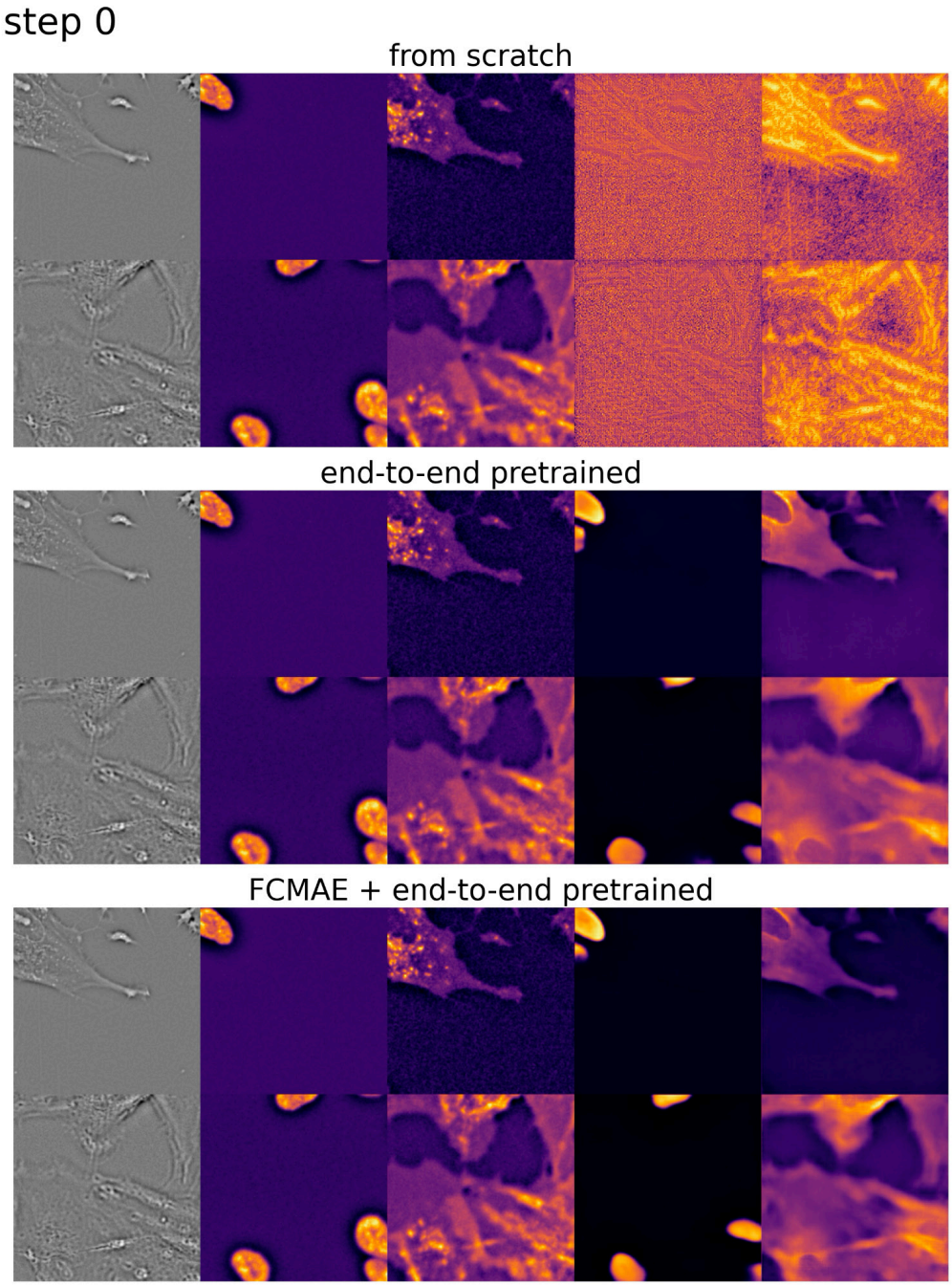
Comparison of neuromast nuclei and membrane fluorescence density and virtual staining 3D segmentations over time. The video shows the 3D instance segmentations at t=0 using the fine-tuned Cellpose model for nuclei and membrane respectively applied to both the fluorescence density and virtual stained and plays over time at the middle z-plane of the neuromast. Virtual staining rescues the uneven expression of nuclei and segments allowing for better segmentation (Scale bar= 25µm)

Video 5



Phase time-lapse images virtually stained with VSCyto2D. Blue: nuclei; Orange: membrane. Every 30 minutes for 24 hours. Scale bar: 100 μm.

Video 6



Validation VSCyto2D predictions during fine-tuning on BJ-5ta cells. Each step is 4 training epochs. Left to right: Phase input patch, nuclei fluorescence (Hoechst), membrane fluorescence (CellMask), nuclei prediction, membrane prediction. Pre-trained models start to produce correct predictions faster. Each image patch is 83.2 µm by 83.2 µm (256 pixels by 256 pixels).