`AnyStar`: Domain randomized universal star-convex 3D instance segmentation

Neel Dey[1*]   S. Mazdak Abulnaga[1]   Benjamin Billot[1]   Esra Abaci Turk[2]

P. Ellen Grant[2]   Adrian V. Dalca[1,3]   Polina Golland[1]

[1]MIT CSAIL   [2]Boston Children's Hospital, Harvard Medical School   [3]Martinos Center, Massachusetts General Hospital

Abstract

Star-convex shapes arise across bio-microscopy and radiology in the form of nuclei, nodules, metastases, and other units. Existing instance segmentation networks for such structures train on densely labeled instances for each dataset, which requires substantial and often impractical manual annotation effort. Further, significant reengineering or finetuning is needed when presented with new datasets and imaging modalities due to changes in contrast, shape, orientation, resolution, and density. We present AnyStar, *a domain-randomized generative model that simulates synthetic training data of blob-like objects with randomized appearance, environments, and imaging physics to train general-purpose star-convex instance segmentation networks. As a result, networks trained using our generative model do not require annotated images from unseen datasets. A single network trained on our synthesized data accurately 3D segments C. elegans and P. dumerilii nuclei in fluorescence microscopy, mouse cortical nuclei in* μCT, *zebrafish brain nuclei in EM, and placental cotyledons in human fetal MRI, all without any retraining, finetuning, transfer learning, or domain adaptation. Code is available at* https://github.com/neel-dey/AnyStar.

1. Introduction

Motivation. Assigning dense semantic labels for biomedical segmentation is difficult, individual instance annotations are even more expensive, and doing so in 3D is often infeasible. Even if a given dataset is painstakingly annotated, any subsequently trained segmentation network is unlikely to generalize to new datasets, scanners, and imaging configurations. For example, an instance segmentation network trained to segment *C. elegans* nuclei in fluorescence microscopy is unlikely to also segment similarly-shaped nuclei in the mouse brain in micro-CT images. Consequently, high-throughput morphometric workflows across biosciences and radiology are bottlenecked by the need for

---

*Webpage: https://www.neeldey.com/any-star/

Table 1. Biomedical instance segmentation methods first train on real image-label pairs and then transfer to each new dataset with or without labels, corresponding to supervised/few-shot learning or domain adaptation, respectively. We instead synthesize image-label training data and generalize without labels or retraining.

Requirements	Supervised Learning	Few-shot Prompting	Domain Adaptation	Ours
Real images for training data	✓	✓	✓	✗
Manual labels for training data	✓	✓	✓	✗
Manual labels for new datasets	✓	✓	✗	✗
(Re-)training on new datasets	✓	✗	✓	✗

reengineering and/or retraining networks for new datasets. Moreover, to retrain or adapt networks, biologists and clinicians need specialized hardware, data annotation pipelines, and machine learning expertise and infrastructure, which discourages rapid data analysis and adoption. We present a *zero-shot* segmentation approach with appropriate appearance & shape priors that addresses practitioner needs.

Current approaches. Major efforts have been made to address these challenges. Domain adaptation methods train on annotated source domain images and either adapt trained networks to unlabeled target domain images [37] or perform cross-domain image translation [10, 40]. Unfortunately, these approaches typically succeed only when source and target domains are closely related [49]. Importantly, they also require biomedical specialists to train unstable and artifact-prone generative models (e.g. CycleGAN [54]) for each new dataset and to produce or acquire structurally similar annotated datasets to use as source domain data.

Another generalist approach collects large-scale inter-modality real-world annotated images for supervised training [38,44]. However, these methods require the new data to be well represented in the training corpus and are currently limited to 2D due to the difficulty of volumetric labeling. To our knowledge, there is no large-scale multi-organism and multi-modality corpus of 3D annotated instances to enable such generalist training with real data. Recent semantic segmentation methods in neuroimaging [4, 5] employ domain randomization [46] and use training labels to simulate synthetic training images to generalize to new modalities and
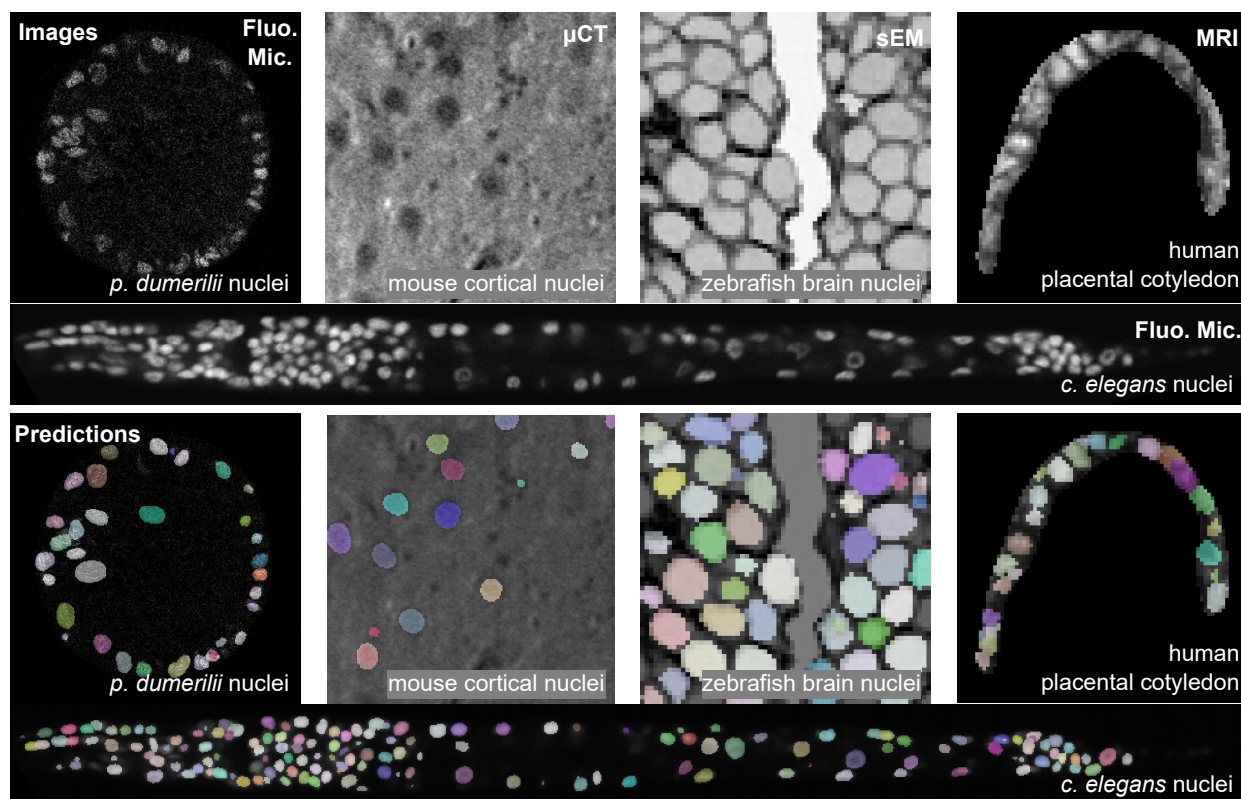
Figure 1. **Qualitative results.** After training an instance segmentation network only on 3D synthetic images generated by the `AnyStar` generative model, a *single* trained network can segment foreground objects in real biomedical images across several imaging modalities, contrasts, and organisms without ever having seen any real images prior to testing and without any retraining or adaptation.

imaging configurations without finetuning in settings with relatively low variation such as anatomical neuroimage segmentation. In contrast, our goal of instance segmentation for several distinct biomedical foreground objects in arbitrary environments across organisms significantly expands image variability and necessitates novel methodology.

Contributions. We present `AnyStar`, a generative model that synthesizes generalist training data to enable instance segmentation networks to segment any star-convex instance across bio-microscopy and radiology (Fig. 1). Biomedical targets such as nuclei and nodules can often be approximated by blob-like star-convex shapes [41, 50]. Using this prior, we simulate image-label training data with domain-randomized object appearances, environments, densities, and imaging physics. An instance segmentation network trained using `AnyStar` gains empirical contrast-invariance in arbitrary environments. As a result, this network generalizes to five completely unseen datasets across biomedical microscopy and radiology without retraining. An `AnyStar`-trained network approaches the segmentation accuracy of fully and weakly supervised domain-specific networks that require expensive annotations and outperforms supervised and/or pretrained networks when they are

presented with out-of-domain data. Finally, we investigate several generative modeling ablations with distinct intensity prior assumptions and find that a single generalist model that incorporates all assumptions is often sufficient.

2. Related work

Biomedical instance segmentation. Established non-deep learning frameworks [8, 28, 35, 43] typically use Otsu [36], feature engineering, and Watershed-based [3] pipelines to segment foreground objects. Using deep networks, early supervised semantic segmentation methods segmented cluttered objects by modeling a boundary class for improved separation [11, 18]. More recently, instance segmentation frameworks such as region-proposing [48,53] and spatial embedding [26] networks typically achieve better instance-specific performance. In particular, [41, 50] found that fitting instances using star-convex shapes led to improved biomedical instance segmentation due to the wide applicability of the star-convex shape prior. Building on the star-convex shape prior, we develop a generative model that removes the need for retraining or adaptation and yields a universal star-convex instance segmentation network.
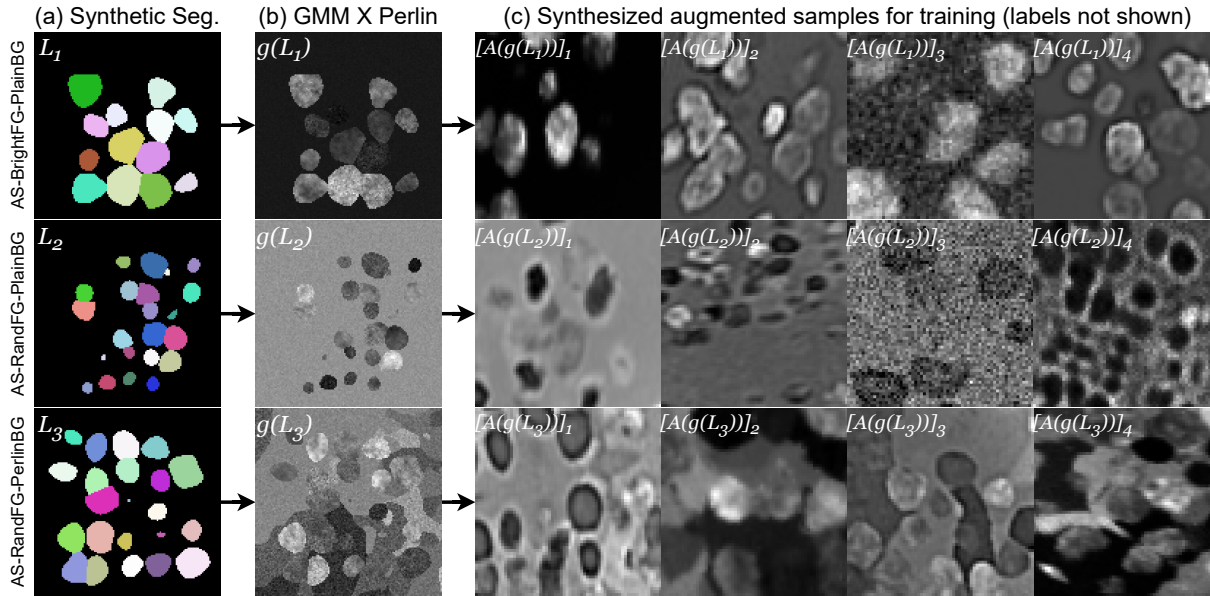
Figure 2. **Generative model.** Starting from n labels in a synthetic segmentation L (**col. 1**), we sample intensities in $g(L)$ from a n-component GMM which is pointwise modulated by Perlin noise (**col. 2**). A carefully designed augmentation sequence $A(\cdot)$ then simulates training data for instance segmentation (**cols. 3–6**). **Rows 1–3** showcase ablations with different specified priors over structure and contrast.

Domain adaptation using generative models. Given an annotated source dataset and a source-domain trained segmentor, cycle-consistent GAN losses are often used to close the domain gap to unlabeled target datasets. While successful in both microscopy [14, 27, 30, 52] and radiology [9,10,23,40], these methods require domain pairs to contain similar structures for stable training and require retraining on each new dataset. In contrast, our method does not require source data (whether annotated or not) for training as it does not need to adapt to new domains at test time.

Generalist models and few-shot prompting. Recently, large models pretrained on multiple real datasets have achieved strong segmentation performance on unseen natural and biomedical 2D datasets using test-time prompts like points/bounding boxes [25] and context sets [6]. 2D models such as Segment Anything [25] can be applied zero-shot slice-wise as well to biomedical images but require finetuning and/or interactive filtering of extraneous predictions for usable performance [12,22,33]. Further, they are practically hard to finetune as 3D biomedical datasets with instance annotations are exceedingly rare and small in sample size, thus hindering their biomedical application. From another perspective, generalist bio-microscopy methods [38,44] merge several 2D datasets to train models on diverse real images with the aim of generalizing to similar images. We instead perform 3D segmentation directly, focus only on star-convex objects, require no interaction or context sets, do not necessitate acquiring multiple real datasets, and generalize to structures not represented in multi-dataset corpora.

Realistic synthetic image generation. Another class of unsupervised methods simulate label maps and train dataset-specific generative models to simulate training images [14–17, 21, 34]. These approaches typically train CycleGAN-like models to translate between the simulated organism-specific labels and real images from the target dataset. They then use the trained generative model to generate image-label pairs to train a second dataset-specific segmentation network. Our approach differs as it synthesizes both images and labels in a domain-agnostic manner and our generative model does not require any domain-specific training.

Biomedical domain randomization. Instead of synthesizing high-fidelity photorealistic training data, domain randomization methods [20, 46, 47] anticipate domain shifts at deployment by synthesizing *unrealistic* training examples with much higher variability using fully-controllable generative models. Consequently, domain-randomized semantic segmentation networks using a small number of training labels (without corresponding real images) [4, 5] achieve strong generalization across modalities and resolutions in neuroimaging. In our work, alongside appearance randomization, AnyStar also performs label randomization to train a universal star-convex instance segmentor for disparate organisms in variable imaging environments.

3. Methods

Star-convexity. A set $S \subset \mathbb{R}^n$ is star-convex if there exists an element $s_0 \in S$ that can be connected to each $s \in S$
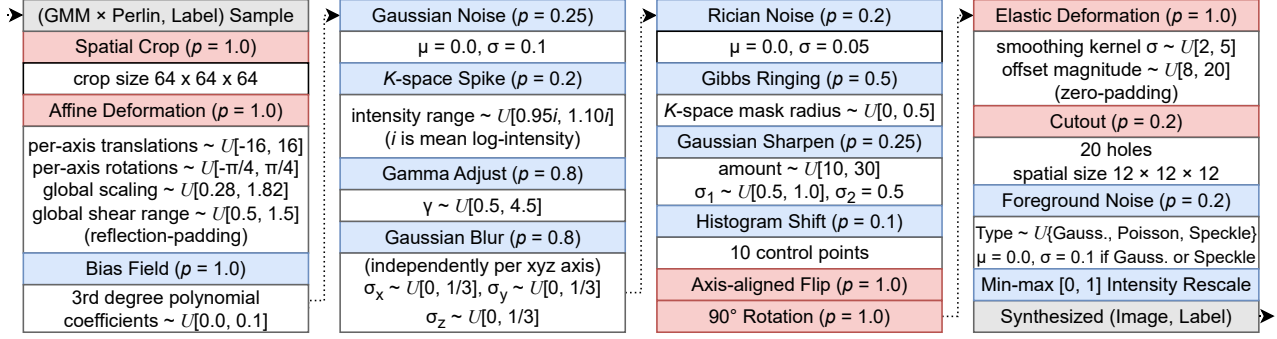
Augmentation pipeline

Column 1:
- (GMM × Perlin, Label) Sample
- Spatial Crop ($p = 1.0$)
- crop size 64 x 64 x 64
- Affine Deformation ($p = 1.0$)
- per-axis translations ~ $U[-16, 16]$ / per-axis rotations ~ $U[-\pi/4, \pi/4]$ / global scaling ~ $U[0.28, 1.82]$ / global shear range ~ $U[0.5, 1.5]$ / (reflection-padding)
- Bias Field ($p = 1.0$)
- 3rd degree polynomial coefficients ~ $U[0.0, 0.1]$

Column 2:
- Gaussian Noise ($p = 0.25$)
- $\mu = 0.0$, $\sigma = 0.1$
- K-space Spike ($p = 0.2$)
- intensity range ~ $U[0.95i, 1.10i]$ (i is mean log-intensity)
- Gamma Adjust ($p = 0.8$)
- γ ~ $U[0.5, 4.5]$
- Gaussian Blur ($p = 0.8$)
- (independently per xyz axis) σ_x ~ $U[0, 1/3]$, σ_y ~ $U[0, 1/3]$ σ_z ~ $U[0, 1/3]$

Column 3:
- Rician Noise ($p = 0.2$)
- $\mu = 0.0$, $\sigma = 0.05$
- Gibbs Ringing ($p = 0.5$)
- K-space mask radius ~ $U[0, 0.5]$
- Gaussian Sharpen ($p = 0.25$)
- amount ~ $U[10, 30]$ σ_1 ~ $U[0.5, 1.0]$, $\sigma_2 = 0.5$
- Histogram Shift ($p = 0.1$)
- 10 control points
- Axis-aligned Flip ($p = 1.0$)
- 90° Rotation ($p = 1.0$)

Column 4:
- Elastic Deformation ($p = 1.0$)
- smoothing kernel σ ~ $U[2, 5]$ offset magnitude ~ $U[8, 20]$ (zero-padding)
- Cutout ($p = 0.2$)
- 20 holes spatial size 12 × 12 × 12
- Foreground Noise ($p = 0.2$)
- Type ~ $U\{$Gauss., Poisson, Speckle$\}$ $\mu = 0.0$, $\sigma = 0.1$ if Gauss. or Speckle
- Min-max [0, 1] Intensity Rescale
- Synthesized (Image, Label)

Figure 3. **Augmentation pipeline** $A(\cdot)$. AnyStar augmentations and their probabilities p and hyperparameters (white boxes), read top-to-down left-to-right. Inputs and outputs are outlined in grey boxes and joint image-label and image-only augmentations are depicted by red and blue boxes, respectively. μ and σ denote means and standard deviations and other notational conventions follow [7].

by a line segment entirely within S. Following [50], we assume 3D biomedical foreground objects to be star-convex polyhedra parameterized by distances to the object boundary along pre-defined unit rays from each internal voxel.

Label synthesis. We first generate synthetic discrete foreground label maps (Fig. 2a) with oblong and irregularly shaped instances by adopting an existing synthetic nuclei generation approach [51]. Specifically, we place n spheres of radius r and centers c_i, $i = \{1, \ldots, n\}$ at the vertices of a regularly spaced 3D grid. The spheres are then independently randomly translated and scaled and up to a third of them are randomly removed. To simulate non-spherical shapes, the distances d_j^i between each voxel j (with coordinates $\mathbf{x}_j$) and object centers $c_1, \ldots, c_n$ are corrupted by voxel-wise additive Perlin noise p_j [39] as $d_j^i = \|\mathbf{x}_j - \mathbf{c}_i\|_2 + 0.9rp_j$. Voxel j is then assigned instance label i if $\min_i d_j^i < r_i$ and is considered background otherwise. These initial label maps are zero- or reflection-padded independently along each axis to simulate varying instance densities and are scaled to a common image grid. We note that other label simulation approaches such as randomly distorted ellipsoids [45] would yield visually similar images using our image synthesis pipeline described below.

Intensity mixture modeling. Given a label map L with n instances, we synthesize an initial image $g(L)$ (Fig. 2b). We sample foreground intensities of $g(L)$ from an n-component Gaussian mixture model (GMM) whose parameters $\{\mu_i, \sigma_i\}_{i=1}^n$ are drawn from a uniform distribution for each image. If a foreground voxel belongs to instance i, then its intensity is sampled from $\mathcal{N}(\mu_i, \sigma_i^2)$. We then apply multiplicative Perlin noise to emulate the spatial texture (e.g., staining differences) common in biomedical imaging.

Background synthesis. To model variable backgrounds and environments in $g(L)$, we investigate several choices corresponding to the rows of Fig. 2. To synthesize bright foreground instances, we model background intensities as

an additional $(n + 1)$th component in the GMM described above with $\mu_{n+1} < \min\{\mu_1, \ldots, \mu_n\}$. Alternatively, to simulate instances that may be brighter or darker than their surroundings, we simply sample $\mu_1, \ldots, \mu_{n+1}$ uniformly at random as before. However, both assumptions do not yet account for instances that may be embedded in strongly textured environments with non-star-convex background structures as in radiology, for example. We therefore build on the shapes generative model of [20] to simulate $b \sim \mathcal{U}\{1, B\}$ random geometric shapes in the background, again using a b-component GMM. To generate spatial background sub-categories to later assign GMM components, we sample a b-channel Perlin noise volume and deform each channel independently with a smooth deformation. We next assign each background voxel a background sub-category corresponding to the channelwise argmax (not used during segmentation training). We then draw from a b-component GMM for each background voxel analogously to the foreground.

Ablations. Our complete generative model, AS-Mix, uses all three of the above background models and randomly assigns one to each synthesized sample. Our generative model ablations include AS-BrightFG-PlainBG and AS-RandFG-PlainBG which simulate bright and randomized contrast foreground instances on untextured background, respectively. AS-RandFG-PerlinBG only uses textured backgrounds with randomized foreground contrast.

Augmentation sequence. L and $g(L)$ are sampled to lie on a 128^3 grid and are augmented by an extensive pipeline $A(\cdot)$ to generate the final training images as illustrated in Fig. 2 cols. 3–6. We randomly crop 64^3 subvolumes, followed by affine spatial deformations (translations, rotations, scales, and shears) with reflection padding used to simulate variable instance densities. We then employ several intensity augmentations including random bias fields, k-space spikes, Gibbs ringing, sharpening, gamma adjustments, and cutout [13] to simulate variable imaging artifacts. Further, Gaussian blurring along each axis independently is used to

Table 2. **Experimental datasets.** 3D datasets used for evaluations. No real volumes are used for `AnyStar` training. Training data listed below are only used for training supervised baselines and validation data are used to tune probability and NMS thresholds only.

Name	Organism	Modality	Image Resolution	Image Grid	#train	#validation	#test
`CE` [19, 31, 32]	*C. elegans*	Fluo. Mic.	$0.116 \times 0.116 \times 0.122 \mu m^3$	$1050 \times 140 \times 140$	18	3	7
`NucMM-Z` [29]	Zebrafish	sEM	$0.480 \times 0.510 \times 0.510 \mu m^3$	$64 \times 64 \times 64$	25	2	27
`NucMM-M` [29]	Mouse	μCT	$0.480 \times 0.510 \times 0.510 \mu m^3$	$192 \times 192 \times 192$	4	-	4
`PlatyISH` [26]	*P. dumerilii*	Fluo. Mic.	$0.450 \times 0.450 \times 0.450 \mu m^3$	$300 \times 300 \times 300$	2	-	1
`Placenta`	Human	BOLD	$3.000 \times 3.000 \times 3.000 mm^3$	$80 \times 80 \times 64$	Qualitative evaluation		

simulate partial voluming common to anisotropic biomedical images. This is followed by spatial deformations using random axis-aligned flips, 90° rotations, and elastic deformations with zero padding. Zero padding and cutout are used at the end of $A(\cdot)$ to simulate blank regions common in bioimaging. Finally, we add Gaussian, Poisson, or speckle noise to the non-zero regions. All augmentations and their stochastic probabilities are summarized in Fig. 3.

Segmentation network. While the training data produced by our generative model can train any 2D or 3D instance segmentation network, we use a `StarDist` network [50] as it matches our expected shape prior. `StarDist` regresses distance maps and "centerness" probability maps whose dense label predictions are filtered by non-maximum suppression to obtain final segmentations. We use its default losses and hyperparameters with 96 rays. As we focus on data generation, we train identical architectures and loss functions for all ablations and upper-bound supervised networks and change only their training and validation data.

Implementation details. The same 5-resolution U-Net architecture (with $2\times$ `Conv-BN-ReLU` blocks) is used for all methods, starting with 32 convolutional channels at the highest resolution and doubling thereafter after each max-pooling for all training runs. All networks are trained for 180,000 iterations using the Adam optimizer [24] with an initial learning rate of 2×10^{-4} which is linearly decayed to 0. Fully and weakly-supervised baseline `StarDist` networks use the same architecture and optimization as `AnyStar`, with their augmentations and training durations adjusted for their optimal performance. For fair comparison, we tune the object detection probability and non-maximum suppression thresholds of `StarDist` for all applicable methods on validation data as in [41, 50]. All `StarDist` networks are implemented using the open-source library: https://github.com/stardist and all augmentation implementations are taken from `MONAI` [7]. Due to `AnyStar`'s extensive augmentation pipeline which can CPU-bottleneck training, we sample hundreds of thousands of augmented synthetic training volumes offline alongside using inexpensive on-the-fly augmentations during training. Further implementation details are provided in the supplementary material.

4. Experiments

Data and preprocessing. Table 2 summarizes the datasets and splits evaluated in this work. Our experimental data includes publicly available annotated datasets (`CE`, `NucMM-Z` & M, `PlatyISH`) and a clinically-acquired fetal EPI MRI time-series dataset (`Placenta`) which is not annotated and thus qualitatively evaluated. These datasets are highly diverse and include a wide variety of object contrasts, densities, and background environments (Fig. 1, top). As `NucMM-M` and `PlatyISH` have limited samples, validation is performed on a held-out crop from the training set.

For the qualitative clinical application, we aim to segment individual placental *cotyledons* in `Placenta` images as their MRI intensities are critical for characterizing fetal oxygenation [1]. Therefore, non-placental tissue is removed from the images using a publicly-available network [2] to focus on placental cotyledons. Further preprocessing details such as image registration, cropping, and resizing are provided in the supplementary material for all datasets.

Baselines. We establish *upper-bounds* on performance using `StarDist` networks trained using all or one of the target dataset's training images, corresponding to fully- and weakly-supervised settings. Weakly-supervised networks are not trained for `NucMM-M` and `PlatyISH` datasets due to their limited sample sizes and supervised networks are not trained for `Placenta` as it does not have training labels. We assess the *out-of-domain* performance of these dataset-specific networks using the other datasets containing similar shapes and contrast. As `AnyStar` does not use real images or annotations, its performance is expected to lie between out-of-domain and fully-supervised baselines.

To test the performance of synthetic domain-randomized training data over collections of large-scale multi-dataset real images, we also compare against the generalist pre-trained `cyto` and `nuclei` models available in Cell-Pose [38]. These models are trained on large-scale publicly available real (non-synthetic) 2D biomedical microscopy images with manual annotations and are used as general-purpose microscopy image segmentors. As suggested, we use their slice-by-slice and along-all-planes 3D segmentation implementation in our experiments, tune the object diameter hyperparameter manually on validation data, and in-
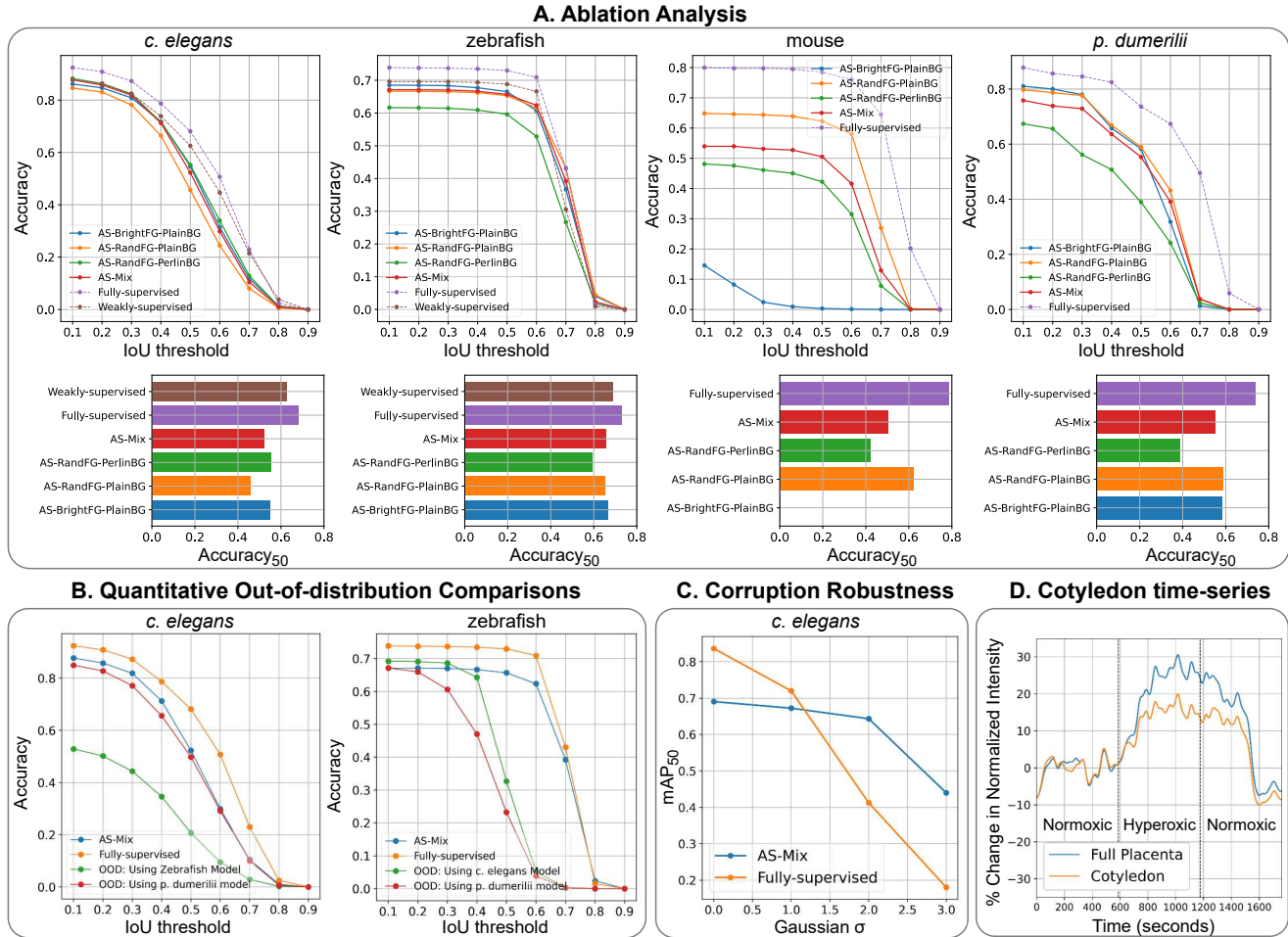
A. Ablation Analysis

c. elegans — zebrafish — mouse — *p. dumerilii*

B. Quantitative Out-of-distribution Comparisons — **C. Corruption Robustness** — **D. Cotyledon time-series**

Figure 4. **Quantitative results. A.** *Top:* Accuracy vs. IoU threshold analysis of the proposed ablations and target dataset-trained fully and weakly-supervised upper bounds across four 3D datasets. *Bottom:* Accuracies at an IoU threshold of 0.5 for easier inspection. **B.** Benchmarking zero-shot segmentation performance against fully-supervised networks trained on similar images from different but structurally-similar datasets. **C.** AnyStar networks are more stable against blur as opposed to domain-specific fully-supervised networks trained with blur augmentation. **D.** Zero-shot instance segmentation enables region-specific temporal analysis of placental MRI.

vert image contrast on the NucMM-M dataset where foreground units have dark contrast. We note that comparisons between our work and CellPose are confounded by changes in architecture and training loss. Lastly, we exclude Segment Anything [25] from comparisons as it is a 2D model, requires significant manual interaction in the form of points or bounding boxes for each 2D slice within large 3D biomedical volumes, and needs users to manually filter multiple extraneous region predictions in its zero-shot mode.

Evaluation criteria. All results are reported on held-out test data. To measure performance, we follow established instance segmentation evaluation strategies. A predicted instance is considered a true positive if it overlaps with a labeled instance by more than a specified IoU threshold. As in [26, 38, 44, 50], to jointly measure detection and segmen-

tation quality, we report mean accuracy computed against increasing the IoU threshold from 0.1 to 0.9 for in/out-of-domain supervised methods and our ablations. Further, as deployed networks may encounter unforeseen image corruptions, we measure the performance change of AS-mix and the fully-supervised network (trained with blur augmentation) when evaluated on images from the highest-resolution dataset (CE) with increasing strengths of Gaussian blurring. We report the mean average precision for these experiments at an IoU threshold of 0.5 with accuracy reported in the supplement. Lastly, previous work on placental oxygenation reports average BOLD intensity across time for the entire placenta which obfuscates differences between distinct subregions [42]. We speculate that cotyledon instance segmentation may reveal biomedically-
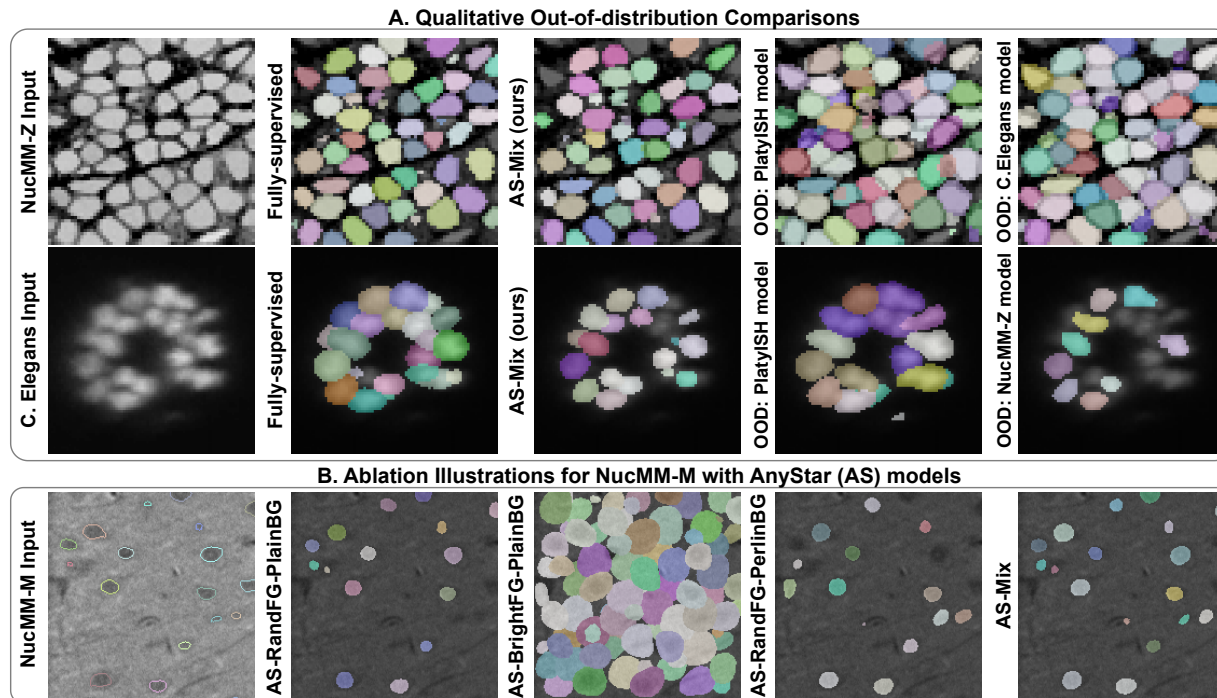
A. Qualitative Out-of-distribution Comparisons

B. Ablation Illustrations for NucMM-M with AnyStar (AS) models

Figure 5. **Out-of-distribution and ablation result visualization**. **A.** Visualizations of zero-shot segmentation performance against fully-supervised networks trained on similar images from other domains. **B.** The outputs of ablation models on a mouse visual cortex sample (`NucMM-M`) with dark foreground (nuclei) contrast in μCT. Ablations using a narrow mis-specified intensity prior (`AS-BrightFG-PlainBG`) do not generalize across organisms and illustrate the importance of randomizing object contrast.

relevant cotyledon intensity time series within placentae.

4.1. Results

Fig. 1 visualizes zero-shot segmentation predictions using `AS-mix` on all five datasets. Figs. 4 and 6A report relevant quantitative statistics. We observe the following:

AS-mix achieves strong inter-dataset generalization. While a specific ablation may outperform others on a specific dataset, usually when its priors match the target dataset appearance, `AS-Mix` achieves consistently strong performance across all datasets (Fig. 4A). The `AS-BrightFG-PlainBG` ablation which trains only on images with bright foreground objects predictably fails when the target dataset (`NucMM-M`) has foreground instances with dark contrast as visualized in Fig. 5B. Consequently, we recommend `AS-Mix` over other ablations for general use, unless the target dataset has a consistent intensity pattern matching one of our other ablations.

AnyStar outperforms pretrained networks on unseen datasets. Fully supervised methods provide an upper bound for in-domain performance when large sets of annotated images are available and task-specific models can be trained. Surprisingly, `AnyStar` approaches the performance of in-domain supervised methods without any in-task training

data. Importantly, relative to supervised networks evaluated on unseen datasets containing instances of similar size, appearance, and contrast, `AnyStar` demonstrates consistently better performance (Figs. 4B, 5A). Representing a mild domain shift, a supervised network trained with extensive augmentation on `PlatyISH` nuclei in fluorescence microscopy does not generalize as well to `CE` nuclei as `AnyStar`, which has never seen this modality. When evaluated on a larger domain gap via the `NucMM-Z` dataset, the fluorescence microscopy models trained on instances with similar contrasts underperform, whereas `AnyStar` produces better segmentations (Figs. 4B, right; 5A, top).

In Fig. 6, in comparison to CellPose's pretrained `cyto` and `nuclei` models, we find two distinct outcomes. On `CE` and `PlatyISH` whose imaging modality (fluorescence microscopy) and shapes (nuclei) are well represented in the CellPose training data, `AS-Mix` performs similarly to these pretrained models (Fig. 6A, left col.). However, on datasets containing imaging modalities unseen by CellPose (`NucMM-Z` & `M`), `AS-Mix` generalizes better, highlighting the benefit of training on randomized synthetic appearances.

AnyStar gains robustness to blur degradation (**Fig. 4C**). Compared to a network trained with all available `CE` training data and augmentation (including blur), `AnyStar-mix` better maintains segmentation performance

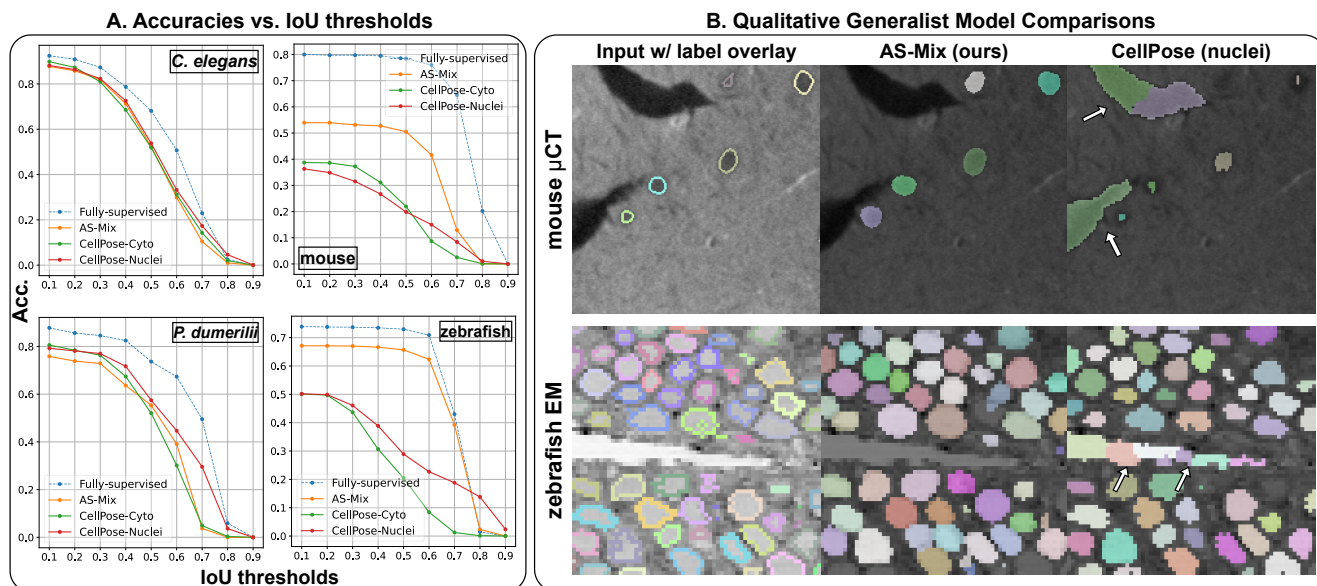Figure 6. **Comparison against generalist models. A.** Accuracy vs. IoU threshold analysis of our model and two pretrained generalist CellPose models trained on a large multi-dataset corpus of real images [38, 44]. **B.** Arbitrarily selected segmentation examples from the `NucMM-M` and `NucMM-Z` datasets produced by our model and CellPose. White arrows point to false positive predictions.

as test images are progressively corrupted by Gaussian filtering, indicating improved robustness to texture changes.

`AnyStar` enables exploratory data analysis on unlabeled data. We select an arbitrary `Placenta` subject and compare the EPI time course over 321 motion-stabilized frames for the entire placenta vs. the average temporal intensities of the centroids of the segmented cotyledon regions. We smooth both their normalized intensities by a temporal Gaussian kernel with $\sigma = 3.0$. In Fig. 4D, using `AS-mix` we visualize the relative BOLD intensity changes in cotyledon subregions within the placenta with maternal oxygenation which showcases `AnyStar`'s practical utility on unlabeled datasets for exploratory data analysis. Further cotyledon segmentations are provided in the supplement.

5. Discussion and conclusions

Limitations and future work. The presented method has limitations which will be addressed in future work. For example, by definition, contrast invariant `AnyStar` networks will segment any star-convex object. This property may yield task-irrelevant predictions on datasets containing multi-contrast star-convex instances with only a few contrasts being of interest. However, task-specific prediction filtering methods based on intensity and shape priors would directly address these applications. Also, as `AnyStar` does not currently represent non-star-convex objects such as neurons and vessels, multiple other shape priors can be integrated into future pipelines. Further, due to strong variability in biomedical image scales and the lack of a canonical resolution (as in neuroimaging [5]), we use sliding window inference on large volumes that performs best when the patches contain objects that are sized similarly to the simulated objects. We expect that multi-scale training and inference methods would overcome this limitation. Importantly, dataset-specific (re-)training typically improves performance over zero-shot methods. If retraining expertise and infrastructure are available in biomedical centers, `AnyStar` can produce zero-shot segmentations that can be quickly refined to construct training sets, thereby reducing annotation effort and enabling rapid prototyping.

Conclusion. In contrast to proposing a new architecture or loss, we focused on synthesizing data to train a generalist biomedical instance segmentation model. To that end, we developed `AnyStar`, a domain-randomized generative model with a carefully designed stochastic appearance and shape model to simulate variable biomedical environments. A *single* network trained on the synthesized data zero-shot segmented 3D biomedical objects across five unseen radiology and bioimaging datasets without any form of retraining and yielded both strong evaluation performance relative to current generalist approaches and enabled a novel and previously infeasible clinical application in fetal MRI.

References

[1] Esra Abaci Turk, Jeffrey N Stout, Christopher Ha, Jie Luo, Borjan Gagoski, Filiz Yetisir, Polina Golland, Lawrence L Wald, Elfar Adalsteinsson, Julian N Robinson, et al. Placental mri: developing accurate quantitative measures of oxygenation. *Topics in magnetic resonance imaging: TMRI*, 28(5):285, 2019. 5

[2] S Mazdak Abulnaga, Sean I Young, Katherine Hobgood, Eileen Pan, Clinton J Wang, P Ellen Grant, Esra Abaci Turk, and Polina Golland. Automatic segmentation of the placenta in bold mri time series. In *Perinatal, Preterm and Paediatric Image Analysis: 7th International Workshop, PIPPI*, 2022. 5

[3] Serge Beucher. Watersheds of functions and picture segmentation. In *ICASSP'82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 7, pages 1928–1931. IEEE, 1982. 2

[4] Benjamin Billot, You Colin, Magdamo Cheng, Sudeshna Das, and Juan Eugenio Iglesias. Robust machine learning segmentation for large-scale analysis of heterogeneous clinical brain MRI datasets. *Proceedings of the National Academy of Sciences (PNAS)*, 120(9):1–10, 2023. 1, 3

[5] Benjamin Billot, Douglas N. Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V. Dalca, and Juan Eugenio Iglesias. Synthseg: Segmentation of brain mri scans of any contrast and resolution without retraining. *Medical Image Analysis*, 86:102789, 2023. 1, 3, 8

[6] Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Universeg: Universal medical image segmentation. *arXiv preprint arXiv:2304.06131*, 2023. 3

[7] M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022. 4, 5

[8] Anne E Carpenter, Thouis R Jones, Michael R Lamprecht, Colin Clarke, et al. Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. *Genome biology*, 7:1–11, 2006. 2

[9] Agisilaos Chartsias, Giorgos Papanastasiou, Chengjia Wang, Scott Semple, David E Newby, Rohan Dharmakumar, and Sotirios A Tsaftaris. Disentangle, align and fuse for multimodal and semi-supervised image segmentation. *IEEE transactions on medical imaging*, 40(3):781–792, 2020. 3

[10] Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng Ann Heng. Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. *IEEE transactions on medical imaging*, 39(7):2494–2505, 2020. 1, 3

[11] Hao Chen, Xiaojuan Qi, Lequan Yu, and Pheng-Ann Heng. Dcan: deep contour-aware networks for accurate gland segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2487–2496, 2016. 2

[12] Ruining Deng, Can Cui, Quan Liu, Tianyuan Yao, Lucas W Remedios, Shunxing Bao, Bennett A Landman, Lee E Wheless, Lori A Coburn, Keith T Wilson, et al. Segment anything model (sam) for digital pathology: Assess zero-shot segmentation on whole slide imaging. *arXiv preprint arXiv:2304.04155*, 2023. 3

[13] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 4

[14] Kenneth W Dunn, Chichen Fu, David Joon Ho, Soonam Lee, Shuo Han, Paul Salama, and Edward J Delp. Deepsynth: Three-dimensional nuclear segmentation of biological images using neural networks trained with synthetic data. *Scientific reports*, 9(1):1–15, 2019. 3

[15] Dennis Eschweiler, Malte Rethwisch, Mareike Jarchow, Simon Koppers, and Johannes Stegmaier. 3d fluorescence microscopy data synthesis for segmentation and benchmarking. *Plos one*, 16(12):e0260509, 2021. 3

[16] Dennis Eschweiler and Johannes Stegmaier. Denoising diffusion probabilistic models for generation of realistic fully-annotated microscopy image data sets. *arXiv preprint arXiv:2301.10227*, 2023. 3

[17] Chichen Fu, Soonam Lee, David Joon Ho, Shuo Han, Paul Salama, Kenneth W Dunn, and Edward J Delp. Three dimensional fluorescence microscopy image synthesis and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2221–2229, 2018. 3

[18] Fidel A Guerrero-Pena, Pedro D Marrero Fernandez, Tsang Ing Ren, Mary Yui, Ellen Rothenberg, and Alexandre Cunha. Multiclass weighted loss for instance segmentation of cluttered cells. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2451–2455. IEEE, 2018. 2

[19] Peter Hirsch and Dagmar Kainmueller. An auxiliary task for learning nuclei segmentation in 3d microscopy images. In *Medical Imaging with Deep Learning*, pages 304–321. PMLR, 2020. 5

[20] Malte Hoffmann, Benjamin Billot, Douglas N Greve, Juan Eugenio Iglesias, Bruce Fischl, and Adrian V Dalca. Synthmorph: learning contrast-invariant registration without acquired images. *IEEE transactions on medical imaging*, 41(3):543–558, 2021. 3, 4

[21] Le Hou, Ayush Agarwal, Dimitris Samaras, Tahsin M Kurc, Rajarsi R Gupta, and Joel H Saltz. Robust histopathology image analysis: To label or to synthesize? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8533–8542, 2019. 3

[22] Yuhao Huang, Xin Yang, Lian Liu, Han Zhou, Ao Chang, Xinrui Zhou, Rusi Chen, Junxuan Yu, Jiongquan Chen, Chaoyu Chen, et al. Segment anything model for medical images? *arXiv preprint arXiv:2304.14660*, 2023. 3

[23] Yuankai Huo, Zhoubing Xu, Hyeonsoo Moon, Shunxing Bao, et al. SynSeg-Net: Synthetic Segmentation Without Target Modality Ground Truth. *IEEE Transactions on Medical Imaging*, 38(4), 2019. 3

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 3, 6

[26] Manan Lalit, Pavel Tomancak, and Florian Jug. Embedseg: Embedding-based instance segmentation for biomedical microscopy data. *Medical Image Analysis*, 81:102523, 2022. 2, 5, 6

[27] Leander Lauenburg, Zudi Lin, Ruihan Zhang, Márcia dos Santos, Siyu Huang, Ignacio Arganda-Carreras, Edward S Boyden, Hanspeter Pfister, and Donglai Wei. Instance segmentation of unlabeled modalities via cyclic segmentation gan. *arXiv preprint arXiv:2204.03082*, 2022. 3

[28] David Legland, Ignacio Arganda-Carreras, and Philippe Andrey. Morpholibj: integrated library and plugins for mathematical morphology with imagej. *Bioinformatics*, 32(22):3532–3534, 2016. 2

[29] Zudi Lin, Donglai Wei, Mariela D Petkova, Yuelong Wu, et al. Nucmm dataset: 3d neuronal nuclei instance segmentation at sub-cubic millimeter scale. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021*, pages 164–174. Springer, 2021. 5

[30] Dongnan Liu, Donghao Zhang, Yang Song, Fan Zhang, Lauren O'Donnell, Heng Huang, Mei Chen, and Weidong Cai. Unsupervised instance segmentation in microscopy images via panoptic domain adaptation and task re-weighting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4243–4252, 2020. 3

[31] Fuhui Long, Hanchuan Peng, Xiao Liu, Stuart K Kim, and Eugene Myers. A 3d digital atlas of c. elegans and its application to single-cell analyses. *Nature methods*, 6(9):667–672, 2009. 5

[32] Fuhui Long, Hanchuan Peng, Xiao Liu, Stuart K Kim, Eugene Myers, Dagmar Kainmüller, and Martin Weigert. 3D nuclei instance segmentation dataset of fluorescence microscopy volumes of C. elegans. *Zenodo*, Feb. 2022. 5

[33] Jun Ma and Bo Wang. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023. 3

[34] Faisal Mahmood, Daniel Borders, Richard J Chen, Gregory N McKay, Kevan J Salimian, Alexander Baras, and Nicholas J Durr. Deep adversarial training for multi-organ nuclei segmentation in histopathology images. *IEEE transactions on medical imaging*, 39(11):3257–3267, 2019. 3

[35] Claire McQuin, Allen Goodman, Vasiliy Chernyshev, Lee Kamentsky, et al. Cellprofiler 3.0: Next-generation image processing for biology. *PLoS biology*, 16(7):e2005970, 2018. 2

[36] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979. 2

[37] Cheng Ouyang, Konstantinos Kamnitsas, Carlo Biffi, Jinming Duan, and Daniel Rueckert. Data efficient unsupervised domain adaptation for cross-modality image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI*, 2019. 1

[38] Marius Pachitariu and Carsen Stringer. Cellpose 2.0: how to train your own model. *Nature Methods*, pages 1–8, 2022. 1, 3, 5, 6, 8

[39] Ken Perlin. An image synthesizer. *ACM Siggraph Computer Graphics*, 19(3):287–296, 1985. 4

[40] M. Ren, N. Dey, J. Fishbaugh, and G. Gerig. Segmentation-renormalized deep feature modulation for unpaired image harmonization. *IEEE Transactions on Medical Imaging*, 2021. 1, 3

[41] Uwe Schmidt, Martin Weigert, Coleman Broaddus, and Gene Myers. Cell detection with star-convex polygons. In *Medical Image Computing and Computer Assisted Intervention–MICCAI*, 2018. 2, 5

[42] Marianne Sinding, David A Peters, Sofie S Poulsen, Jens B Frøkjær, Ole B Christiansen, Astrid Petersen, Niels Uldbjerg, and Anne Sørensen. Placental baseline conditions modulate the hyperoxic bold-mri response. *Placenta*, 61:17–23, 2018. 6

[43] Christoph Sommer, Christoph Straehle, Ullrich Koethe, and Fred A Hamprecht. Ilastik: Interactive learning and segmentation toolkit. In *2011 IEEE international symposium on biomedical imaging: From nano to macro*, pages 230–233. IEEE, 2011. 2

[44] Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: a generalist algorithm for cellular segmentation. *Nature methods*, 18(1):100–106, 2021. 1, 3, 6, 8

[45] David Svoboda, Michal Kozubek, and Stanislav Stejskal. Generation of digital phantoms of cell nuclei and simulation of image formation in 3d image cytometry. *Cytometry Part A: The Journal of the International Society for Advancement of Cytometry*, 75(6):494–509, 2009. 4

[46] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017. 1, 3

[47] Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 969–977, 2018. 3

[48] Eric Upschulte, Stefan Harmeling, Katrin Amunts, and Timo Dickscheid. Contour proposal networks for biomedical instance segmentation. *Medical image analysis*, 77:102371, 2022. 2

[49] Gijs van Tulder and Marleen de Bruijne. Unpaired, unsupervised domain adaptation assumes your domains are already similar. *Medical Image Analysis*, page 102825, 2023. 1

[50] Martin Weigert, Uwe Schmidt, Robert Haase, Ko Sugawara, and Gene Myers. Star-convex polyhedra for 3d object detection and segmentation in microscopy. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3666–3673, 2020. 2, 4, 5, 6

[51] Martin A. Weigert and Uwe Schmidt. Stardist 3d example notebook. 4

[52] Liming Wu, Alain Chen, Paul Salama, Kenneth Dunn, and Edward Delp. Nisnet3d: Three-dimensional nuclear synthesis and instance segmentation for fluorescence microscopy images. *bioRxiv*, 2022. 3

[53] Zhuo Zhao, Lin Yang, Hao Zheng, Ian H Guldner, Siyuan Zhang, and Danny Z Chen. Deep learning based instance segmentation in 3d biomedical images using weak annotation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI*, 2018. 2

[54] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 1