

---

# WEAK-TO-STRONG GENERALIZATION ENABLES FULLY AUTOMATED DE NOVO TRAINING OF MULTI-HEAD MASK-RCNN MODEL FOR SEGMENTING DENSELY OVERLAPPING CELL NUCLEI IN MULTIPLEX WHOLE-SLICE BRAIN IMAGES

---

A PREPRINT

Lin Bai<sup>1</sup> Xiaoyang Li<sup>1,4</sup> Liqiang Huang<sup>1</sup>  
 Quynh Nguyen<sup>1</sup> Hien Van Nguyen<sup>1</sup> Saurabh Prasad<sup>1</sup>  
 Dragan Maric<sup>2</sup> John Redell<sup>3</sup> Pramod Dash<sup>3</sup>  
 Badrinath Roysam<sup>1\*</sup>

<sup>1</sup> Department of Electrical and Computer Engineering, University of Houston, TX 77024, USA

<sup>2</sup> National Institute of Neurological Disorders and Stroke, Bethesda, MD 20892, USA

<sup>3</sup> Department of Neurobiology and Anatomy, UT McGovern Medical School, Houston, TX, USA

<sup>4</sup> Canva Pty Ltd., Sydney, NSW, Australia

\*Corresponding author: broysam@uh.edu

December 11, 2025

## ABSTRACT

We present a weak to strong generalization methodology for fully automated training of a multi-head extension of the Mask-RCNN method with efficient channel attention for reliable segmentation of overlapping cell nuclei in multiplex cyclic immunofluorescent (IF) whole-slide images (WSI), and present evidence for pseudo-label correction and coverage expansion, the key phenomena underlying weak to strong generalization. This method can learn to segment *de novo* a new class of images from a new instrument and/or a new imaging protocol without the need for human annotations. We also present metrics for automated self-diagnosis of segmentation quality in production environments, where human visual proofreading of massive WSI images is unaffordable. Our method was benchmarked against five current widely used methods and showed a significant improvement. The code, sample WSI images, and high-resolution segmentation results are provided in open form for community adoption and adaptation.

**Keywords** weak to strong learning · automated training · instance segmentation · whole-slide imaging · cell nuclei · multiplex immunofluorescence

## 1 Introduction

Multiplex immunofluorescent (IF) imaging of whole slides of tissues using automated cyclic staining is currently performed on a large scale for deep cellular characterization of histological samples in pre-clinical science and drug discovery ([Al-Kofahi et al., 2009]; [Li et al., 2017]; [Lin et al., 2003]; [Lin et al., 2007]; [Lin et al., 2005]; [Maric et al., 2021]; [Rivest et al., 2023]). These systems generate massive multi-gigabyte multi-channel whole-slide images (WSI) containing hundreds of thousands or more cells that must be analyzed accurately with maximum-possible and scalable automation. Reliable detection and accurate segmentation of cell nuclei is a critical yet challenging first step towards cell-based quantification of molecular markers. This problem is especially challenging since WSI images cover extended tissue regions that can exhibit high variability in staining, tissue architecture, and cell morphology, and varied

image artifacts arising from automated scanning and cyclic immunofluorescence processes. Interestingly, even the fluorescent staining of cell nuclei is highly variable across WSI tissue regions and the commonly used DNA stain DAPI does not mark all nuclei as evident in **Figure 1A** showing an image of a rat brain slice ( $29,398 \times 43,054$  pixels per channel,  $>250,000$  cells) drawn from a 50-plex dataset ([Maric et al., 2021]) (The full high-resolution image is posted in the Supplement).

Recent progress in segmenting cell nuclei has largely been driven by the application of deep neural networks and foundational models trained on massive human-annotated datasets ([Kirillov et al., 2023]; [Radford et al., 2021]; [Ramesh et al., 2022]) in a bid to capture the variability. Despite the progress, these methods remain human annotation effort intensive. Foundational models require massive initial annotation efforts to train, and additional annotation efforts to adapt the models to a new class of images. Even then, several problems remain. For example, the correct handling of overlapping nuclei in thicker samples remains unaddressed (**Figure 2**). Next, foundational models are often trained on 3-channel (RGB) images and therefore cannot exploit the cues that are available in multiplex WSI images, a missed opportunity (**Figure 1**). Finally, the problem of assessing the accuracy of automatically generated segmentations remains human effort intensive. There is a compelling need for automated assessment methods, especially in production environments, where visual proofreading of massive WSI images is unaffordable and not scalable.

We present an integrated *fully automated* methodology based on *weak-to-strong generalization* ([Lang et al., 2024]) to overcome the above-mentioned limitations (**Figure 3**), and metrics for automated segmentation quality assessment. Weak-to-strong generalization is a special case of weakly supervised machine learning in which a strong model (student) is trained using (comparatively unreliable) annotations generated automatically by a weaker model (teacher) after they are subjected to well-designed automated cleanup and data augmentation steps. This strong model is trained using weak annotations generated using a current foundational model (*e.g.*, SAM ([Kirillov et al., 2023])) that lacks needed capabilities like the ability to handle overlapping nuclei. In this work, we refer to these weak annotations as “pseudo labels” in line with the weak-to-strong generalization literature. A key step is the automated processing of the pseudo labels to generate a large collection of synthetic examples with numerous instances of complex and variable nuclear overlap patterns. From these augmented weak label data, we train the strong model by weak-to-strong learning ([Lang et al., 2024]). Lang *et al.* ([Lang et al., 2024]) showed that a well-trained student model can learn to correct the weaker model’s mistakes (“pseudo label correction”) and generalize to instances where the teacher lacks confidence, even when these examples are excluded from the training data (“coverage expansion”). We present evidence for these phenomena and show their beneficial impact. Finally, we present metrics for assessing the accuracy of large-scale nuclear segmentations in a fully automated manner based on accounting for the fluorescence signal and quantifying the purity of fluorescent signatures over cells.

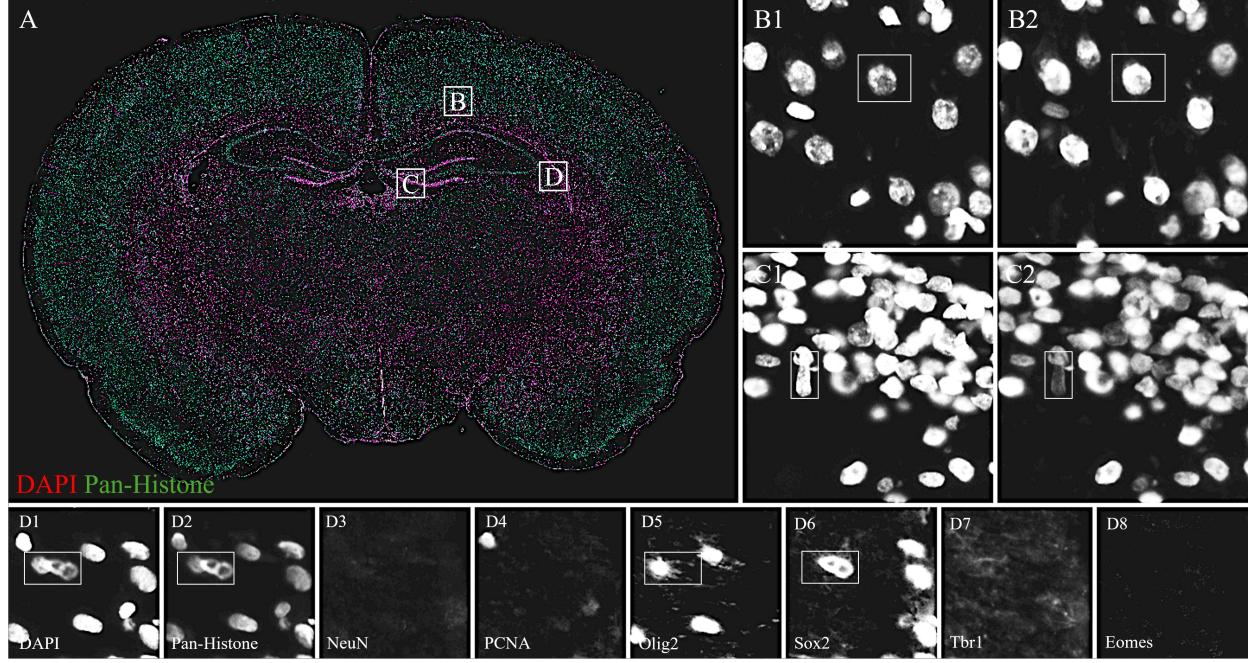


Figure 1: Sample 8-pixel image of a whole rat brain slice ( $29,398 \times 43,054$  pixels per channel,  $>250,000$  cells) showing major spatial variations in DAPI and histone labeling. (A) RGB rendering with DAPI in Red, Pan-Histone in Green, and their average in blue. (B1, B2) Close-up of boxed region B showing strong histone labeling but weak DAPI labeling. (C1, C2) Close-up of boxed region C showing strong DAPI labeling but weak histone labeling. (D1–D8) Close-ups of boxed region D in which we see a cluster of 3 cells that are difficult to separate based on nuclear stains alone that become possible to distinguish when cell-type channels Olig2 and Sox2 are utilized.

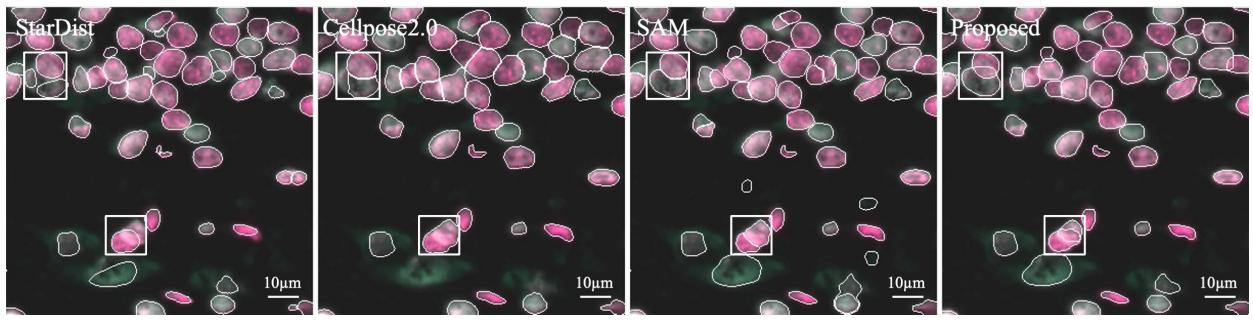


Figure 2: **Figure 2.** Current methods (StarDist, Cellpose, and SAM) exhibit poor performance in segmenting overlapping/clustered nuclei. For example, they either over segment (StarDist) or under segment (Cellpose) or only able to segment the clearly visible part of nuclei body (SAM). However, these methods can be used to generate automated weak annotations to train our strong model that produces accurate results without requiring any human annotation effort.

## 2 Summary of prior work

Methods for segmenting cell nuclei fall into three main categories. First, there are parameter-based methods, such as thresholding ([Otsu, 1975]), watershed ([Malpica et al., 1997]), or active contours ([Chan and Vese, 2001]), that require users to manually adjust parameters for each imaging modality and scale. This requires care and expertise and often yields suboptimal performance, largely due to variability across the large WSI images. The second involves deep CNN models (e.g., Mask-RCNN([He et al., 2017]), U-Net ([Ronneberger et al., 2015]), YOLACT ([Bolya et al., 2019])) that are custom trained for each class of images to cope with the expected variability. Variations of these models ([Graham et al., 2019]; [Jia et al., 2021]; [Johnson, 2018], [Johnson, 2020]; [Long, 2020]; [Lv et al., 2019]; [Vuola et al., 2019]; [Zhou et al., 2018]) have been extensively applied to nuclear segmentation. Deep Learning Models ([Ke et al., 2021]; [Kong et al., 2020]; [Molnar et al., 2016]; [Roy et al., 2023]; [Zhou et al., 2019]) like DoNet ([Jiang et al., 2023]) are specifically designed to identify overlapping objects, and many augmentation methods ([Dvornik et al., 2018]; [Dwibedi et al., 2017]; [Fang et al., 2019]; [Ghiasi et al., 2021]; [Lin et al., 2022]; [Yu et al., 2023]; [Yun et al., 2019]; [Zhang, 2017]) are proposed to boost segmentation performance. While these methods achieve high accuracy, they require extensive annotation efforts, making them time-consuming and resource intensive. The third and most scalable approach is to develop reusable deep learning based models like StarDist ([Schmidt et al., 2018]), Cellpose ([Pachitariu and Stringer, 2022]; [Stringer and Pachitariu, 2025]; [Stringer et al., 2021]), CellSeg ([Lee et al., 2022]) that are pre-trained on large and diverse datasets. While these models are efficient and widely used, they often suffer from domain shift – when the test images arise from a distribution that is different from the training set, and the model’s performance can degrade. Furthermore, these methods are designed to work with a small number of channels (e.g., RGB images), making them sub-optimal in multiplex WSI applications.

Recently, large-scale, pre-trained foundation models (*e.g.* the Segment Anything Model (SAM) ([Kirillov et al., 2023])) trained on >1B masks over 11M images show competitive segmentation performance and often outperform fully-supervised models. Foundation models can generalize data distributions beyond those encountered during training, a capability facilitated by prompt engineering and referred to as zero-shot and few-shot generalization. Although powerful, models like SAM cannot handle overlapping nuclei, and multiplex images since they are trained on RGB images, and require manual prompts to generate segmentations, limiting their applicability in fully automated workflows. Many methods to fine tune SAM for biomedical images have been described, such as All-in-SAM ([Cui et al., 2024]), fine-tuning SAM by providing box level annotations,  $\mu$ SAM ([Archit et al., 2025]), fine-tuning SAM for light and electron microscopy, MedSAM ([Ma et al., 2024]), fine-tuning SAM on an unprecedented dataset with more than one million medical image-mask pairs. Unfortunately, fine-tuning SAM requires a large corpus of (expensive) domain-specific annotation data.

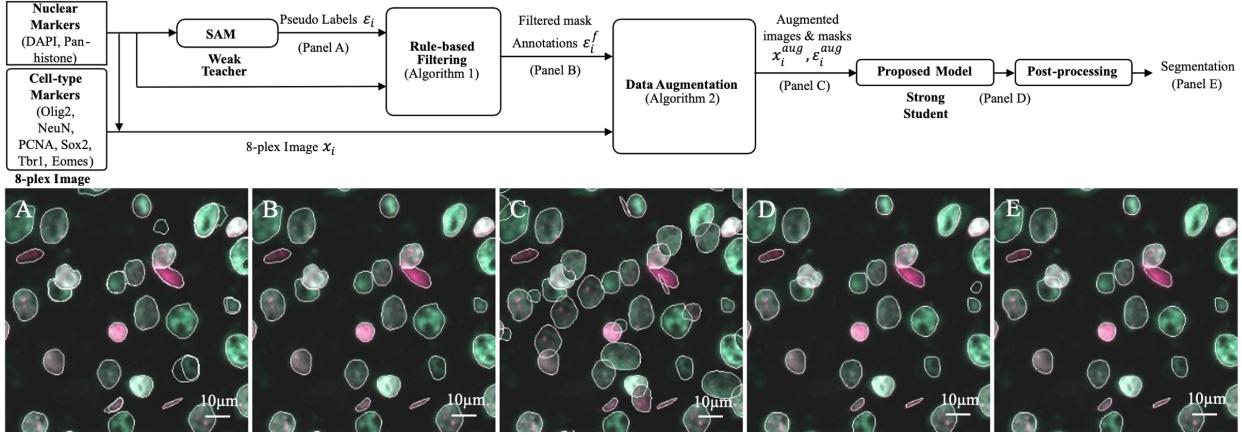


Figure 3: Illustrating our weak-to-strong learning framework, where the foundation model SAM serves as a weak teacher to generate annotations, and our proposed model acts as a strong student to learn from the generated annotations. (A) First, DAPI and Pan-Histone images, along with automated point prompts, are input into SAM (weak teacher) to produce mask annotations. (B) A rule-based filtering process is then applied to remove union masks and duplicate masks, ensuring high-quality annotations. (C) The filtered mask annotations, combined with 8-pixel images, are then passed through a data augmentation module, performing instance-level augmentation. (D) Finally, the augmented data is used to train our proposed model (strong student) for improved segmentation post-processing is applied to produce (E).

In this paper we show how we can leverage existing models (*e.g.*, SAM) as a teacher model to generate pseudo labels for training a more capable student model. In weakly supervised learning, for instance, co-teaching ([Han et al., 2018]), involves training two network to teach the other. Co-teaching+ ([Yu et al., 2019]) builds on this strategy by selecting low-loss points from samples where the two networks disagree. Co-learning ([Tan et al., 2021]) integrates supervised and self-supervised learning with a shared feature encoder with two exclusive heads, that provide different views of the data. JoCoR ([Zhang et al., 2018]) aims to reduce the diversity of two networks during training. Compared to classification, fewer studies focus on segmentation tasks. ADELE ([Liu et al., 2022]) employs an early stopping strategy to allow the network to fit clean labels before it memorizes false annotations. Recently, some authors have developed theoretic support for weak-to-strong learning ([Lang et al., 2024]) for classification tasks, and identified the conditions under which this strategy is successful. Our work aims to leverage these emerging insights and adapt weak to strong learning to instance segmentation.

### 3 Weak To Strong Learning Method

**Figure 3** depicts the main steps of the proposed weak-to-strong learning approach. For simplicity, we describe the detailed steps for fully automated training of a strong model given a single multiplex image  $I$ . This procedure is easily extended to a batch of images. The trained model can be used to segment similar other images (in inference mode).

#### 3.1 Automated Pseudo-label Generation using the Segment Anything Model

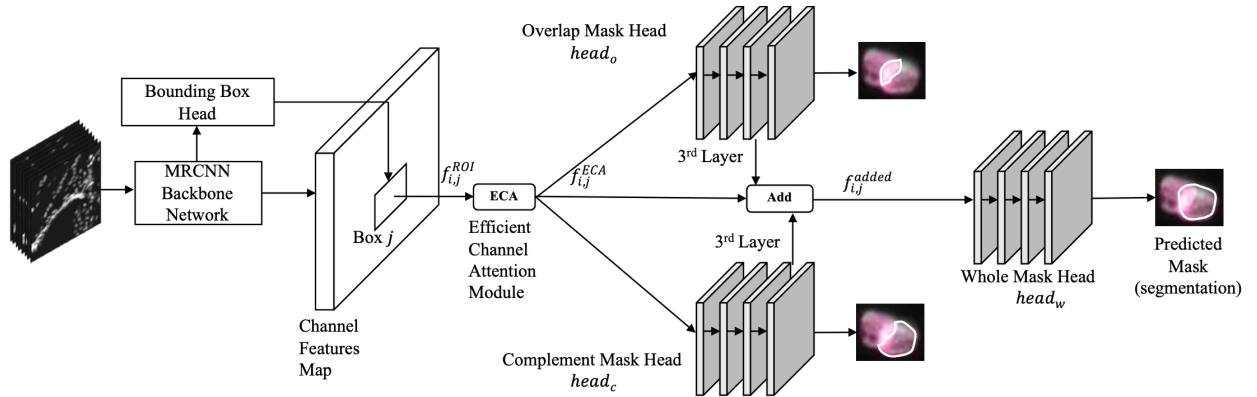


Figure 4: *Overview of the proposed M3-RCNN model architecture. At the detection stage, DAPI and Pan-Histone channels are used as inputs to generate bounding boxes. At the segmentation stage, all eight channels are utilized. For each detected box region, the corresponding channel features are processed through the Efficient Channel Attention (ECA) module, which dynamically determines the importance of each channel's contribution to the output. The weighted features are then passed through a series of convolutional layers to adjust their feature representations. The adjusted features are subsequently fed into three segmentation heads: the Overlap Mask Head, Complement Mask Head, and Whole Mask Head, each predicting different components of the final mask. The intermediate outputs from the Intersection Mask Head and Complement Mask Head are added as input to the Whole Mask Head for enhanced feature learning.*

Given a large multiplex WSI image  $I$ , we dice it into a set  $\{x_i\}_{i=1}^K$  of  $K$   $512 \times 512$ -pixel tiles. We extract a second set of two-channel tiles  $\{x_i^{DAPI, Histone}\}_{i=1}^K$  composed of the nuclear markers DAPI and Pan-histone. From this set, we construct a dataset  $D = \{(x_i^{DAPI, Histone}, \varepsilon_i)\}_{i=1}^K$  where  $\varepsilon_i = \{e_{i,j}\}_{j=1}^{N_i}$  is the set of masks (pseudo labels) generated by SAM using an array of point prompts,  $N_i$  is the number of masks in the  $i^{th}$  tile, and  $e_{i,j}$  is the  $j^{th}$  mask in the  $i^{th}$  tile. We also construct a multiplex version of the dataset  $D = \{(x_i, \varepsilon_i)\}_{i=1}^K$ . At this stage, SAM does not identify the categories of the detected objects (*e.g.*, nuclei/other). The masks  $\varepsilon_i$  may include noisy labels due to false detections, over segmentations, under segmentations, or erroneous overlaps. Accurately identifying these errors requires human proofreading. However, we can filter the masks considerably using a simple rule-based strategy focused on two main types of errors: union masks (two or more smaller masks that are subsumed by a larger mask), and duplicate masks. The output of the rule-based filtering is denoted  $D^f = \{(x_i^{DAPI, Histone}, \varepsilon_i^f)\}_{i=1}^K$ , where  $\varepsilon_i^f = \{e_{i,j}^f\}_{j=1}^{N_i^f}$ ,

and  $N_i^f$  indicates the revised count of instances in the  $i^{th}$  image. Details of the rule-based filtering are presented in pseudocode form as **Algorithm 1** in the Appendix.

### 3.2 Data Augmentation for Simulating Cell Overlaps

To generate a set of training examples of cell overlaps, we use a straightforward ‘copy and paste’ strategy. Within each image  $x_i$ , we select two nuclei from  $\varepsilon_i^f$ , one to serve as the ‘copy’ nucleus, which undergoes a combination of spatial and intensity transformations, and the other as the ‘paste’ nucleus, onto which the first nucleus is overlayed. Adjustments are made to the annotations to reflect these changes. Given that  $\varepsilon_i^f$  may still include noisy labels, we select the ‘copy’ and ‘paste’ nuclei based on four criteria: (1) choosing nuclei that are spatially isolated from others; (2) avoiding nuclei close to the image boundaries; (3) excluding nuclei with concave mask contours; and (4) discarding dim nuclei. For the transformation of ‘copy’ nuclei, we apply random rotation and opacity adjustments to the selected nuclei to simulate expected overlap patterns. To prevent extreme overlaps, we implement a threshold on the overlap ratio. The output of these augmentations is denoted  $D^{aug} = \{(x_i^{aug}, \varepsilon_i^{aug})\}_{i=1}^K$ , where  $\varepsilon_i^{aug} = \{e_{i,j}^{aug}\}_{j=1}^{N_i^{aug}}$ ,  $N_i^{aug}$  represents the revised instance count in the  $i^{th}$  image. A pseudo-code description is provided as **Algorithm 2** in the Appendix.

### 3.3 The Proposed Strong Model ( $M^3$ -RCNN)

**Figure 4** depicts the architecture of the proposed strong  $M^3$ -RCNN model. We use the Mask-RCNN ([He et al., 2017]) as the starting point given its competitive performance in instance segmentation. It operates in two stages. The first stage employs a backbone network like ResNet-50 ([He et al., 2016]) and a feature pyramid network (FPN) to extract multiscale features of each image tile  $x^i$ . A region proposal network (RPN) identifies boxed regions of interest (ROI), and their features  $f_{i,j}^{ROI}$ . The second stage consists of a bounding box regression head, classification head and a mask head that uses the ROI features  $f_{i,j}^{ROI}$  to generate bounding boxes, object classifications and segmentation masks. The composite loss function of Mask-RCNN is articulated as:

$$L_{MRCNN} = \lambda_1 L_{reg} + \lambda_2 L_{cls} + \lambda_3 L_{mask}. \quad (1)$$

where  $L_{reg}$  represents the smooth-L1 loss for bounding box regression in two stages,  $L_{cls}$  is the cross-entropy loss for classification in two stages,  $L_{mask}$  is the pixel-wise cross-entropy loss for segmentation in stage two, and  $\lambda_1$ ,  $\lambda_2$ ,  $\lambda_3$  are loss weights. Here, given that we are analyzing eight-channel images, we use an Efficient Channel Attention (ECA) module ([Wang et al., 2020]) to weight each channel’s feature map to compute the overall feature map  $f_{i,j}^{ECA}$  for each box. The ECA module consists of three components, global average pooling (GAP), one-dimensional convolution ( $W$ ), and a sigmoid function  $\sigma$ . The input to the ECM is  $f_{i,j}^{ROI}$  for the  $j^{th}$  object from  $i^{th}$  image tile, and the output  $f_{i,j}^{ECA}$  is given by:

$$f_{i,j}^{ECA} = \sigma(W * GAP(f_{i,j}^{ROI})). \quad (2)$$

Since nuclei can often overlap, traditional amodal segmentation approaches concentrate on the visible segments of objects, largely due to the inherent challenge of representing occluded areas where they are not directly observable. However, in our context, the objects are not fully occluded, and some visual cues of the occluded nuclei remain visible. To leverage these weak but important cues, we diverge from the conventional single-mask-head framework and introduce two additional mask heads specifically designed to capture the overlapping and complement parts of the objects, alongside the standard whole object representation, as illustrated in **Figure 4**. We denote these mask heads  $head_c$ ,  $head_w$ ,  $head_o$ , for the complement, whole mask, and overlap mask heads, respectively. As shown in **Figure 4**, the inputs of  $head_c$  and  $head_o$  are  $f_{i,j}^{ECA}$ , and the output feature maps of the  $head_c$  and  $head_o$  from the 3<sup>rd</sup> convolution layer are added with  $f_{i,j}^{ECA}$  to obtain the feature map  $f_{i,j}^{added}$  of the whole mask head. The loss functions for the mask heads are formulated as follows:

$$L_{mask}^{nonconcave} = \frac{1}{K} \sum_{i=1}^K \frac{1}{N_i^{aug,nc}} \sum_{j=1}^{N_i^{aug,nc}} (L_w + L_c + L_o), \quad (3)$$

$$L_w = L_{ce}(head_w(f_{i,j}^{added}), e_{i,j}^{aug}), \quad (4)$$

$$L_o = L_{ce}(head_o(f_{i,j}^{ECA}), e_{i,j}^{aug,o}), \quad (5)$$

$$L_c = L_{ce}(head_c(f_{i,j}^{ECA}), e_{i,j}^{aug,c}). \quad (6)$$

where  $L_{mask}^{non-concave}$  is the loss for non-concave samples and  $N_i^{aug,nc}$  represents the total number of instances with non-concave contours,  $e_{i,j}^{aug,o}$  and  $e_{i,j}^{aug,c}$  represent the overlapping and complement parts of the augmented mask  $e_{i,j}^{aug}$ , respectively, and  $L_{ce}$  is the cross-entropy loss function. Given two overlapping nuclei A and B, the segmentation from SAM generally yield a comprehensive mask for nucleus A and a complementary mask for nucleus B. When analyzing its segmentation output, denoted as  $e_{i,j}^{aug}$  and armed with the prior understanding that the typical shape of nuclei is non-concave, we categorize these masks into concave (caused by overlapping) and non-concave based on the morphology of their contours. Masks exhibiting non-concave outlines are considered accurate and can be accepted as is. Masks with concave or incomplete outlines are excluded from the mask loss computation due to their structural irregularity. However, these instances are still utilized in the classification and bounding box regression losses, as their bounding boxes can remain accurate when overlap is not severe, despite the mask being incomplete. Our M<sup>3</sup>-RCNN model is trained in an end-to-end manner using a composite loss function which is given by:

$$LM3RCNN = \lambda_1 L_{reg} + \lambda_2 L_{cls} + \lambda_3 L_{mask}^{nonconcave}. \quad (7)$$

## 4 Experimental Results

The materials and methods for the multiplex WSI imaging are described in our earlier paper and the data are publicly available ([Maric et al., 2021]). We evaluated the proposed method against five current methods (**Table 1**, **Table 2**) in two different ways: (i) automated (unsupervised) assessment using novel metrics described below; and (ii) manual (supervised) assessment against ground truth annotations. The natural near symmetry of the rat brain provides an efficient and practical way to perform these steps. First, our proposed M<sup>3</sup>-RCNN model was trained on  $K = 2,494$  image tiles of size  $512 \times 512$  pixels cropped from the left side of the first whole rat brain slice image (S1). The unsupervised assessment was carried out over all the brain slices (S1, ..., S4), by dividing the WSI images into tiles of size  $512 \times 512$  pixels with 50 pixels overlap. For the more effort-intensive supervised assessment, we randomly selected 100 tiles from the right side of the brain in the S1 dataset, annotated them using the semi-automated online computer vision annotation tool (CVAT), and used this as the ground truth dataset.

Finally, to show the pseudo-label correction phenomenon under our weak-to-strong generalization framework, we randomly selected 100 tiles from our training set (left side of the S1 brain), annotated them using the CVAT annotation tool, and visualized this phenomenon using t-SNE multi-dimensional projection ([Van der Maaten and Hinton, 2008]).

For supervised evaluation, we use Aggregated Jaccard Index plus (AJI+) ([Graham et al., 2019]) and Panoptic Quality (PQ) ([Kirillov et al., 2019]) as our primary metrics. AJI+ improves upon the original Aggregated Jaccard Index (AJI) ([Kumar et al., 2017]) by penalizing unmatched ground truth instances. PQ jointly evaluate detection and segmentation quality in a single metric.

### 4.1 Automated (Unsupervised) Segmentation Quality Assessment Metrics

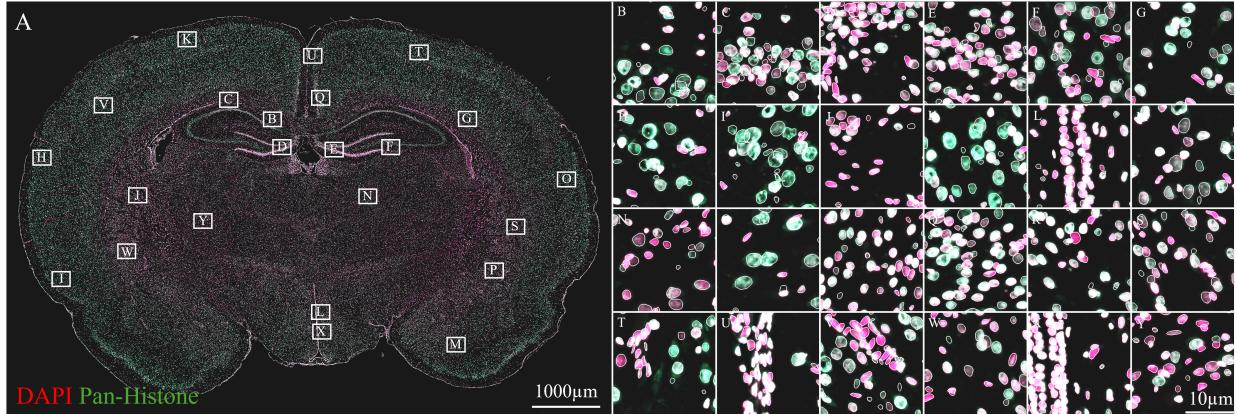


Figure 5: Close-ups of the proposed model’s output across the brain slice, demonstrating its ability to segment overlapping nuclei despite the variability. The complete high-resolution segmentation files for four slices S1 – S4 are provided in the electronic supplement.

As noted earlier, our goal is to assess segmentation quality in a fully automated manner, without the need for human-generated ground truth annotations. We recognize that this goal cannot be met for instance segmentation in general. However, in the context of immunofluorescence imaging, especially multiplex imaging, it is indeed possible to develop practically useful quality metrics by examining the extent to which the fluorescence signal is accounted for by the segmentation masks (“signal coverage”), and the extent to which each object represents an individual cell of a certain type as indicated by the expression of a cell-type marker (we call that marker purity). These metrics are described below.

#### 4.1.1 (1) Signal Coverage Metric $\Gamma$

This is a pixel-level measure that quantifies the number of foreground pixels that are missed by all the segmentation masks generated by the proposed model, collectively denoted  $M_{model}$ . For this, we identify all the foreground pixels  $M_{DAPI, Histone}$  by fitting a two-component Gaussian Mixture Model (GMM) to separate foreground and background. With this, the coverage metric  $\Gamma \in (0, 1)$  over the entire image is formulated as:

$$\Gamma = \frac{\sum_{(x,y)} (M_{model} \cap M_{DAPI, Histone})}{\sum_{(x,y)} M_{DAPI, Histone}}. \quad (8)$$

where  $(x, y)$  denotes all the pixel locations.

#### 4.1.2 (2) Marker Purity Metric $\Pi$

This measure analyzes the expression of cell-type protein markers over each object, as illustrated in **Figure 6**. We consider an object to be ‘pure’ if it contains foreground pixels for one cell type alone, and ‘impure’ otherwise. To make this notion precise, we model a nuclear mask  $M_n$  as the union of multiple cell-type-specific masks  $M_n^c$ , with weights  $\alpha_n^c \in (0, 1)$ , where  $c = 1, \dots, C$  indexes the cell-type marker channels, and  $\alpha_n = [\alpha_n^1, \dots, \alpha_n^C]$  is the vector of these weights. We perform a sparse decomposition using Orthogonal Matching Pursuit (OMP) ([Pati et al., 1993]) to find the most significant cell-type masks and use the corresponding coefficients  $\alpha_n^c$  to define the purity metric for a single mask  $M_n$  as follows.

$$\hat{\alpha}_n = \underset{\alpha_n}{\operatorname{argmin}} \left\| \text{ReLUOMP} \left( M_n - \sum_{c=1}^C \alpha_n^c M_n^c \right) \right\|_2^2, \quad \text{s.t. } \|\alpha_n\|_0 \leq k. \quad (9)$$

where  $k$  is the  $L_0$  sparsity constraint (we use  $k = 3$ ), and ReLUOMP is the rectified linear unit function.

With this, the purity metric for the  $n^{th}$  object is given by  $\Pi_n = \max(\alpha_n^1, \dots, \alpha_n^C)$ , and the global purity metric  $\Pi \in (0, 1)$  over all  $N$  detected masks is given by:

$$\Pi = \frac{1}{N} \sum_{n=1}^N \Pi_n. \quad (10)$$

The object-wise purity metric  $\Pi_n$  can also be used to identify erroneously segmented objects for corrective editing if desired. For the cell-type-specific masks  $M_n^c$ , we use a GMM model to identify the foreground pixels in the cell-type marker channel. In our case, the biomarkers were NeuN for neurons, Iba-1 for microglia, Olig2 for oligodendrocytes, S100

ensuremath

beta for astrocytes and GFP for endothelial cells. Specifically, we first fit a GMM with three components and then identify the second largest and largest GMM clusters based on their mean intensity values. We then calculate the maximum intensity within the second-largest cluster and the minimum intensity within the largest cluster. The average of these two values is used as the foreground threshold. This method is simple yet effective, and filters out the cytoplasm, processes and auto-fluorescence signals.

## 4.2 Performance Summary

**Table 1** and **Table 2** provide a performance summary of the proposed  $M^3$ -RCNN model against five benchmarks (*Cellpose2.0*, *StarDist*, *SAM-C*, *MRCNN*, *MRCNN-AUG*). For the *Cellpose2.0* and *StarDist* models, we used their default parameter settings. For the *SAM-C* method, we provided a  $256 \times 256$  array of point prompts to *SAM* as input

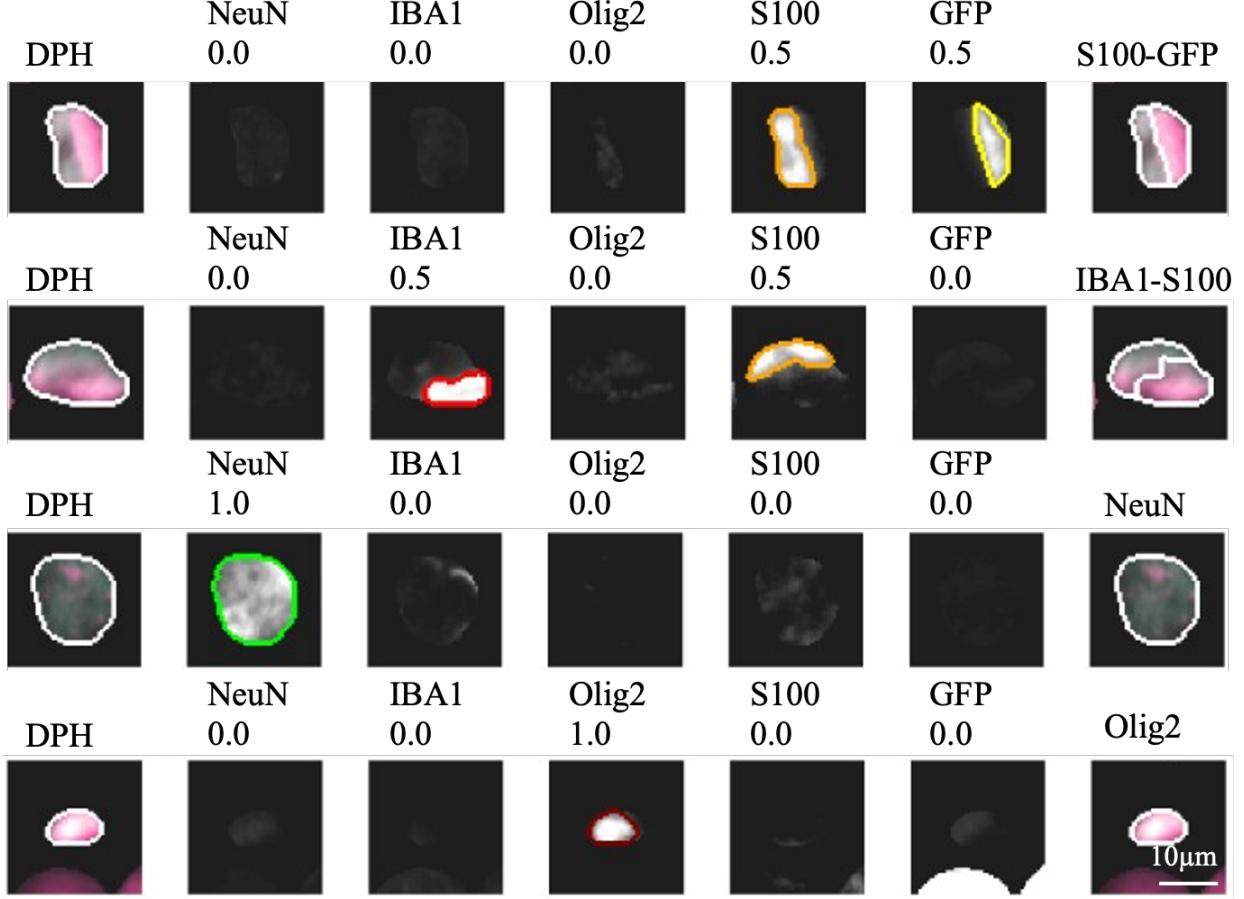


Figure 6: Four examples illustrate the sparse decomposition method for measuring marker purity for cells. DPH represents DAPI and Pan-Histone. The first column shows the nuclear segmentations, and the second to sixth columns show the segmentations of signals in five different channels containing cell-type markers. The last column shows the final type. The first row shows an example of a cell with low purity (0.5) since it contains a mixture of S100 (0.5) and GFP (0.5) signals. The second row shows an example of a cell with low purity (0.5) since it contains a mixture of IBA1 (0.5) and S100 (0.5) signals. The third and fourth rows show high-purity (1.0) examples.

and used our rule-based filtering **Algorithm 1** to eliminate minor errors. For *MRCNN*, we trained the *MRCNN* model on annotations generated by *SAM*, which were filtered using **Algorithm 1**. The *MRCNN-AUG* model was trained on the annotations generated by *SAM*, which were filtered using **Algorithm 1**, and in addition, we applied data augmentation using **Algorithm 2**.

**Table 1:** Performance Metrics on the S1 Dataset

Methods	Channels	S1 AJI+	PQ	Coverage	Purity	Cell Counts
<i>Cellpose2.0</i>	1-plex	65.6%	65.2%	82.89%	98.99%	193,658
<i>StarDist</i>	1-plex	67.4%	68.1%	88.47%	99.19%	247,103
<i>SAM-C</i>	2-plex	76.0%	74.7%	97.62%	99.12%	257,860
<i>MRCNN</i>	2-plex	76.3%	76.0%	98.04%	<b>99.30%</b>	<b>260,267</b>
<i>MRCNN-AUG</i>	2-plex	76.2%	75.9%	<b>98.26%</b>	99.17%	258,782
<i>M3-RCNN</i>	8-plex	<b>77.3%</b>	<b>76.7%</b>	98.22%	99.19%	255,726

**Table 2:** Performance Metrics on the S2, S3, S4 Datasets

Methods	Channels	S2		S3		S4		Cell Counts
		Coverage	Purity	Cell Counts	Coverage	Purity	Cell Counts	
<i>Cellpose2.0</i>	1-plex	74.56%	98.57%	175,823	84.11%	99.11%	197,136	81.17% 98.60% 193,264
<i>StarDist</i>	1-plex	93.81%	98.94%	<b>250,311</b>	88.66%	99.40%	240,768	88.11% 99.11% 243,169
<i>SAM-C</i>	2-plex	91.88%	98.88%	248,263	97.65%	99.33%	<b>266,559</b>	97.40% 99.02% <b>272,346</b>
<i>MRCNN</i>	2-plex	93.12%	<b>99.00%</b>	237,956	99.05%	<b>99.40%</b>	252,179	98.57% <b>99.11%</b> 256,516
<i>MRCNN-AUG</i>	2-plex	<b>94.26%</b>	98.93%	243,830	<b>99.44%</b>	99.36%	257,147	<b>99.00%</b> 99.03% 260,994
<i>M3-RCNN</i>	8-plex	93.93%	98.90%	241,796	98.39%	99.39%	255,823	98.61% 99.01% 256,789

1-plex: DAPI, 2-plex: DAPI, Pan-histone, 8-plex: DAPI, Pan-histone, NeuN, Olig2, PCNA, Sox2,

Tbr1, Eomes

*SAM-C*: SAM-Cleaned (Teacher), *MRCNN-AUG*: Mask-RCNN + Augmentation

The proposed (strong) model was trained using the same annotations as the *MRCNN-AUG* model. To ensure a fair comparison, we trained *M<sup>3</sup>-RCNN*, *MRCNN*, and *MRCNN-AUG* for the 5000 iterations with a batch size 32,  $\lambda_1 = \lambda_2 = \lambda_3 = 1$ ,  $\beta_1 = 0.8$ ,  $\beta_2 = 0.7$ ,  $\beta_3 = 0.5$ ,  $t = 3$ , utilizing the same backbone ResNet-50, model hyperparameters, and post-processing. Post-processing was performed after supervised evaluation in two stages. First, false positive detections were removed by cross-referencing the results with those from *MRCNN*. Then, potentially missed detections were recovered by incorporating segmentation outputs from *Cellpose2.0* and *StarDist*. For the S1 dataset, our method achieves the highest AJI+ and PQ scores. *MRCNN* and *MRCNN-AUG* both outperform *Cellpose2.0*, *StarDist* and *SAM-C*. In terms of the purity and coverage metrics for the S1 - S4 datasets, all three methods including *M<sup>3</sup>-RCNN*, *MRCNN* and *MRCNN-AUG* outperform *Cellpose2.0*, *StarDist* and *SAM-C*. This trend aligns with the results obtained using supervised metrics, further validating the reliability of our proposed evaluation metric. We can also see that our baseline *SAM-C* has surpassed *StarDist* and *Cellpose2.0* due to its training on a large dataset, giving it a better generalization ability. It is not surprising that *Cellpose2.0* and *StarDist* do not perform as well since they have not been trained on this specific dataset. *M<sup>3</sup>-RCNN*, *MRCNN* and *MRCNN-AUG* outperform *SAM-C* which shows the beneficial effect of the coverage expansion phenomenon in weak to strong generalization.

**Table 3: Summary of Results from the Ablation Study**

Components	S1		
ECA	Multi-heads	AJI+	PQ
✓		77.0%	76.3%
	✓	77.1%	76.5%
✓	✓	<b>77.3%</b>	<b>76.7%</b>

**Table 3** provides an ablation study of the proposed *M<sup>3</sup>-RCNN* model. Removing ECA module or multi-heads causes a decrease in performance, showing that each component contributes independently. The complete model which combines both modules achieves the best performance. As shown in **Figure 5**, our model effectively segments entire nuclei for overlapping cases, even in the densely packed regions. *Our approach can be viewed as a fully automated refinement method, enhancing existing segmentation models without the need for expensive ground truth data for training or fine-tuning*.

#### 4.3 Evidence for Weak to Strong Generalization

**Table 4: Measured Expansion and Error Bounds**

Model	$P(\left(\bar{R}(f)\middle S_i\right))$	$c$	$\alpha_i$	$err((f, \tilde{y} S_i))$	Bound	$err((f, y S_i))$
<i>M3-RCNN</i>	0.15	0.43	0.19	0.17	0.32	0.16

*SAM-C* serves as the weak teacher, and the proposed *M<sup>3</sup>-RCNN* serves as the strong student model in the weak to strong generalization framework. The data augmentation in **Algorithm 2** is a crucial link between the weak teacher and

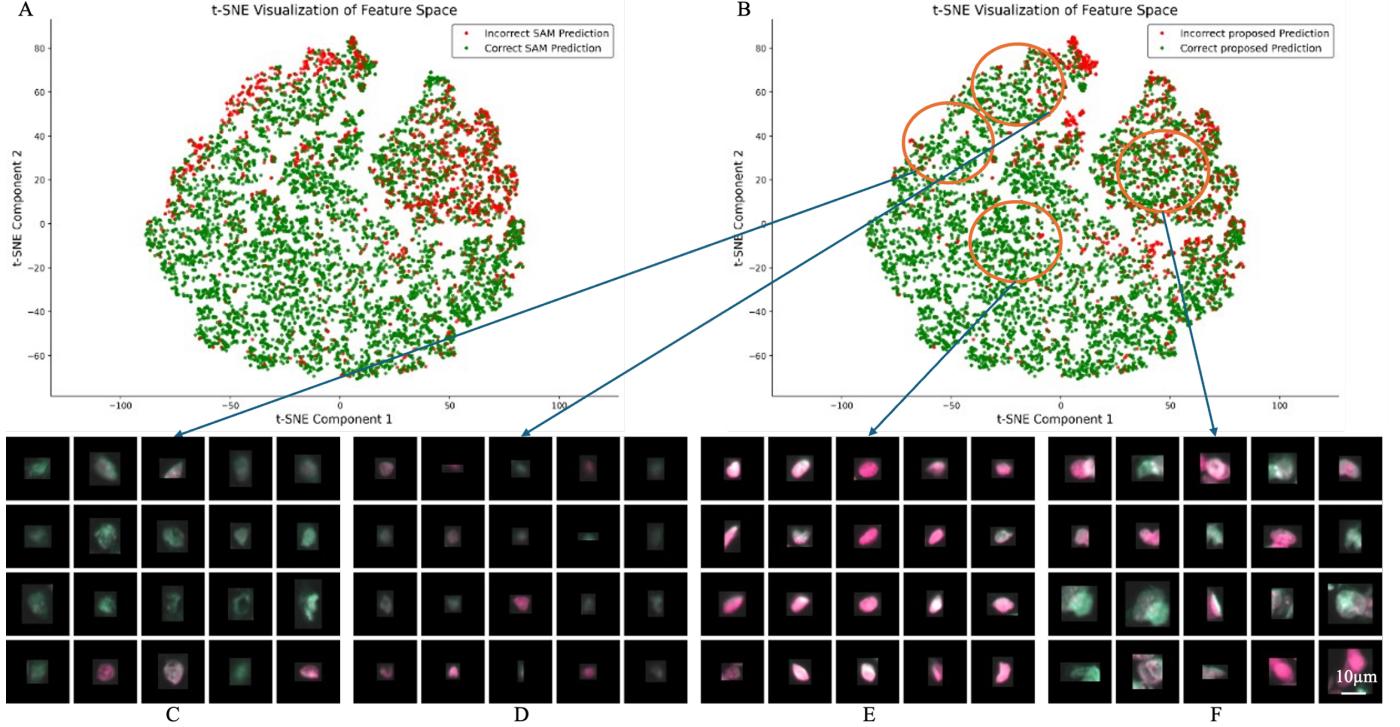


Figure 7: Illustrating the occurrence of pseudo-label correction in feature space. (A) t-SNE plot with each point corresponding to a nucleus in the pseudo-label dataset before the weak-to-strong learning, where red/green indicate incorrect/correct predictions. (B) t-SNE plot after the weak-to-strong learning shows an increase in the number of green data points. By visualizing the clusters of data points (C–F), we observe improved segmentation of dim nuclei and overlapping nuclei, demonstrating the effectiveness of the learning process.

the strong student model. It is clear from **Table 1** that the  $M^3$ -RCNN strong student model outperforms the teacher model on the 100-tiles ground truth dataset that was unseen by the student model and this is evidence for the coverage expansion phenomenon (generalizing to instances where the teacher either lacks confidence or to unlabeled data points that are not part of the training dataset) we expect in weak to strong generalization.

Next, we compared the pseudo-label data before and after weak-to-strong learning to examine evidence for the pseudo-label correction phenomenon (a well-trained student model can learn to correct the weaker model’s mistakes) in feature space (**Figure 7**). For this, we extracted all nuclei from 100 labeled images from the training set by locating their centers and zero padded to  $80 \times 80$  – pixel image patches. Each such patch (DAPI and Pan-histone) is represented by a seven-dimensional (7-D) feature vector that quantifies segmentation difficulty in terms of: (i) foreground contrast; (ii) occlusion score; (iii) boundary variability; (iv) nucleus size; (v) aspect ratio; (vi) edge intensity; and (vii) background variability. Foreground contrast measures the difference in intensity between the nucleus and its background. Occlusion score evaluates the proportion of a nucleus that is overlapped by other nuclei. Boundary variability evaluates the number of edge pixels relative to nucleus size. Nucleus size normalizes the nucleus area by image size. Aspect Ratio is the width-to-height ratio. Edge intensity is the mean intensity of nucleus edges. Lastly, Background variability measures the standard deviation of background pixel intensities.

We used the t-SNE algorithm to project these 7-D features to the plane for visualization and examined close-ups of image regions for visual confirmation. In the t-SNE projection, a green dot corresponds to a correctly predicted nucleus, and a red dot corresponds to an incorrectly predicted nucleus. To determine whether a prediction is correct (green) or incorrect (red), we used the following criteria. For non-overlapping nuclei, a prediction is considered correct if its IoU with the ground truth exceeds 0.7. For overlapping nuclei, a prediction must cover at least 10 pixels of the overlapping region and have an  $\text{IoU} > 0.5$  prediction to be considered as correct. In comparing SAM-C and our proposed model (**Figure 7**), we clearly observe a decrease in the total number of red points indicating the pseudo-label correction phenomenon, and a visual confirmation of improved segmentation accuracy. Additionally, by visualizing the clusters of data points, particularly in regions where pseudo-label correction occurred, we see that our model achieves better segmentation of large, small and dim nuclei, as well as nuclei occluded by other nuclei.

Lang *et al.* ([Lang et al., 2024]) have calculated an upper bound for the error rate  $\text{err}(f, y|S_i)$  of a student model against ground truth data on a pseudo labeled training subset  $S_i$  with known ground truth. Their estimate was formulated for classification rather than segmentation problems. With this in mind, we choose to consider our model  $M^3\text{-RCNN}$  as a classifier  $f$  that classifies image patches, denoted  $x$ , that are centered on nuclei. Following the notation of Lang *et al.*, we denote  $y$  as the ground truth labeler, and  $\tilde{y}$  as the pseudo labeler  $SAM\text{-C}$ . With this notation,  $\text{err}(f, y|S_i)$  is the error rate for  $f$  against the ground truth labels, where  $S_i$  represents the set of extracted image patches around the nuclei (**Figure 7C-F**). We use  $S_i^{bad}$  to denote the incorrectly pseudo labeled image patches and  $S_i^{good}$  to denote the correctly pseudo labeled patches. With this, the upper bound for  $\text{err}(f, y|S_i)$  is given by:

$$\text{err}(f, y|S_i) \leq \frac{2\alpha_i}{1 - 2\alpha_i} P(\overline{R(f)}|S_i) \text{err}(f, \tilde{y}|S_i) \alpha_i \left(1 - \frac{3}{2}c\right). \quad (11)$$

where  $\alpha_i = P(S_i^{bad}|S_i)$  is the error rate of the pseudo labeler against ground truth labels, and  $P(\overline{R(f)}|S_i)$  is a measure of the classifier  $f$ 's robustness, where a lower value corresponds to a higher robustness, and *vice versa*. Here,  $R(f) = \{x : r(f, x) = 0\}$ , where  $r(f, x) = P(f(x') \neq f(x)|x' \in \mathcal{N}(x))$  is the probability that  $f$  assigns different labels to  $x$  and its neighbor  $x'$  in feature space. In equation (11),  $c$  is the expansion rate between  $S_i^{bad}$  and  $S_i^{good}$ , which measure how well two sets are mixed together, and is calculated using the same formulation described in Lang *et al.* ([Lang et al., 2024]). The calculated values are shown in **Table 4**. Our model error rate  $\text{err}(f, y|S_i)$  is smaller than  $\alpha_i$ , this indicates the occurrence of the pseudo correction phenomenon. Also, the upper bound of  $\text{err}(f, y|S_i)$  equals to 0.32, which is reasonably close to the actual observed value of 0.17, indicating that the theoretical estimation aligns well with empirical results.

In summary, we see clear evidence of the two key phenomena associated with weak to strong generalization – pseudo-label correction and coverage expansion. Importantly, these phenomena result in a superior nuclear segmentation algorithm.

## 5 Conclusions and Discussion

We have demonstrated the practicality of building strong (more capable) nuclear segmentation models from weak (less capable) models in a fully automated manner under the framework of weak to strong generalization, for which evidence is clearly observed. Importantly, the entire process, starting from the image to segmentation results accompanied by performance metrics, is fully automated and does not require any manual annotation effort. *This strategy is made possible by the availability of versatile and powerful foundational models like SAM, and the exploitation of image cues that are available in multiplex IF whole-slide images.* Data cleanup and augmentation algorithms are the unsung but crucial enabler of weak to strong generalization.

Our method can be used to segment novel image datasets *de novo* (e.g., images from a new instrument using a new protocol) in a fully automated manner. The first run of the proposed method on a new dataset produces two outcomes: (i) accurate fully automated segmentation results for the new dataset accompanied by automated assessment metrics; and (ii) a trained  $M^3\text{-RCNN}$  model that can be used for segmenting (in inference mode) a larger collection of similar datasets (e.g., images from the same instrument using the same protocols). For whole-slide images (WSIs) with dimensions of  $29,398 \times 43,054$  pixels, our pipeline is capable of generating complete image predictions within ~9 hours (1 min to divide WSIs into tiles, 4 hours 25 mins to generate masks using SAM, 11 mins to run **Algorithm 1**, 2 mins to run **Algorithm 2**, 2 hours for training, 1 hour 2 mins for inferencing and postprocessing, 57 mins for merging results) on a computer with six NVIDIA RTX6000A GPUs, fully automatically. Once trained, subsequent inferencing using the model takes ~0.01 seconds per tile in the WSI image with 6,012 tiles (~1 hour total).

By introducing ECA module and multi mask heads, our  $M^3\text{-RCNN}$  model learned capabilities that are not present in the Mask-RCNN model (e.g., improved handling of overlapped nuclei). The proposed automated (unsupervised) evaluation metrics offer the ability to assess automated segmentation quality without the need for human inspection and proofreading. This becomes more valuable in the current context when automated multiplex IF systems are routinely producing massive images on a sufficiently large scale that manual proofreading is unaffordable. When manual editing is needed, the purity metric  $\Pi_n$  can be used to identify the subset of objects that need corrective editing (efficient analytics-driven proofreading).

The proposed model is amenable to further refinement and can be adapted to other image analysis tasks. Going forward, it is possible to develop even stronger models and update the cleanup and augmentation modules in concert to develop novel capabilities, with full automation. Even with brain tissue images, there is an opportunity to refine the method in

extremely dense regions (*e.g.*, the hippocampus). While our data augmentation algorithm cannot fully simulate the full range of nuclear overlapping scenarios in such regions, we note that human analysis of these regions is so challenging that reliable human annotations are difficult to obtain. Ultimately, the proposed method cannot generalize to cases that are entirely absent from the training set.

### Appendix A1: Algorithm 1 (Rule-based filtering)

```

Input:
D = {(x_i^{\{DAPI,Histone\}}, eps_i)}_{i=1..K}, eps_i = {e_{i,j}}_{j=1..N_i}
parameters: beta1, beta2, beta3
Output:
D_filtered = {(x_i^{\{DAPI,Histone\}}, eps_i_filtered)}_{i=1..K}

for i = 1..K do
    eps_i_u_filtered = {}                      # store non-union masks
    for j = 1..N_i do
        eps_i_components = {}                  # store component masks
        for m = 1..N_i do
            if j != m and SumPixels(e_{i,j}) CAP e_{i,m}) / SumPixels(e_{i,m}) >
                beta1 then
                    add e_{i,m} to eps_i_components
            end if
        end for
        if SumPixels(merge(eps_i_components) CAP e_{i,j}) / SumPixels(e_{i,j}) >
            beta2
            and count(eps_i_components) > 1 then
                pass                                # union mask (discard)
            else
                add e_{i,j} to eps_i_u_filtered
                # optional: for each component e_{i,m} in eps_i_components:
                #   if isnuclei(x_i^{\{DAPI,Histone\}}, e_{i,j} - e_{i,m}) then
                #       add (e_{i,j} - e_{i,m}) to eps_i_u_filtered
                #   end if
            end if
        end for
    eps_i_d_filtered = {}                      # store non-duplicate masks
    for p = 1..count(eps_i_u_filtered) do
        for q = 1..count(eps_i_u_filtered) do
            if p != q and
                SumPixels(eps_i_u_filtered[p] CAP eps_i_u_filtered[q]) / SumPixels(
                    eps_i_u_filtered[p]) > beta3 and
                SumPixels(eps_i_u_filtered[p]) / SumPixels(eps_i_u_filtered[q]) < 1
                    then
                pass                                # duplicate (discard)
            else
                add eps_i_u_filtered[p] to eps_i_d_filtered
                # optional:
                #   if isnuclei(x_i^{\{DAPI,Histone\}}, eps_i_u_filtered[p] -
                #       eps_i_u_filtered[q]) then
                #       add (eps_i_u_filtered[p] - eps_i_u_filtered[q]) to
                #       eps_i_d_filtered
                #   end if
            end if
        end for
    end for
    eps_i_filtered = {}
    for r = 1..count(eps_i_d_filtered) do
        if notdim(x_i^{\{DAPI,Histone\}}, eps_i_d_filtered[r]) then
            add eps_i_d_filtered[r] to eps_i_filtered
        end if
    end for
end for

```

## Appendix A2: Algorithm 2 (Data augmentation)

```

Input:
D_f = {(x_i, eps_i^f)}_{i=1..K}, eps_i^f = {e_{i,j}^f}_{j=1..N_i^f}
parameter: t
Output:
D_aug = {(x_i^aug, eps_i^aug)}_{i=1..K}

for i = 1..K do
    obj_i_copy = eligible_copy(x_i, eps_i^f)
    obj_i_paste = eligible_paste(x_i, eps_i^f)

    for j = 1..count(obj_i_copy) do
        obj_i_trans = augmentation_transform(obj_i_copy[j])
        (x_i, eps_i^f) = copy_paste(obj_i_trans, obj_i_paste, t, x_i, eps_i^f)
    end for

    x_i^aug = x_i
    eps_i^aug = eps_i^f
end for

```

## A3 - Supplemental Files

The code, four sample WSI images (S1 – S4), and full-resolution segmentation results are provided in open form for viewing, community adoption, and potential adaptation to other biomedical image analysis tasks. The WSI images (S1 – S4) can be found at figshare. The code and results are shared in Dropbox.

## Acknowledgments

This work was supported by the National Institutes of Health grant R01NS109118.

## References

- Yousef Al-Kofahi, Wiem Lassoued, William Lee, and Badrinath Roysam. Improved automatic detection and segmentation of cell nuclei in histopathology images. *IEEE Transactions on Biomedical Engineering*, 57(4):841–852, 2009.
- Xiaoyang Li, Jokubas Jahanipour, Dragan Maric, and Badrinath Roysam. Computational mapping of rat brain cytoarchitectonics using multiplex biomarker imaging and quantitative analysis. *Unpublished manuscript*, 2017.
- Gang Lin, Umesh Adiga, Karin Olson, John F Guzowski, Carol A Barnes, and Badrinath Roysam. A hybrid 3d watershed algorithm incorporating gradient cues and object models for automatic segmentation of nuclei in confocal image stacks. *Cytometry Part A: the journal of the International Society for Analytical Cytology*, 56(1):23–36, 2003.
- Gang Lin, Monica K Chawla, Karin Olson, Carol A Barnes, John F Guzowski, Christoffer Bjornsson, William Shain, and Badrinath Roysam. A multi-model approach to simultaneous segmentation and classification of heterogeneous populations of cell nuclei in 3d confocal microscope images. *Cytometry Part A: the journal of the International Society for Analytical Cytology*, 71(9):724–736, 2007.
- Gang Lin, Monica K Chawla, Karin Olson, John F Guzowski, Carol A Barnes, and Badrinath Roysam. Hierarchical, model-based merging of multiple fragments for improved three-dimensional segmentation of nuclei. *Cytometry Part A: the journal of the International Society for Analytical Cytology*, 63(1):20–33, 2005.
- Dragan Maric, Jokubas Jahanipour, Xiaoyang R Li, Adarsh Singh, Aryan Mobiny, Hien Van Nguyen, Andrea Sedlock, Kedar Gramma, and Badrinath Roysam. Whole-brain tissue mapping toolkit using large-scale highly multiplexed immunofluorescence imaging and deep neural networks. *Nature communications*, 12(1):1550, 2021.
- François Rivest, Deniz Eroglu, Bjorn Pelz, Julia Kowal, Annika Kehren, Vytautas Navikas, Maria Grazia Procopio, Philippe Bordignon, Emmanuel Peres, Martin Ammann, Elodie Dorel, Sara Scalmazzi, Luigi Bruno, Markus Ruegg, Gael Campargue, Guillermo Casqueiro, Luc Arn, Jonas Fischer, Saska Brajkovic, and Denis Dupouy. Fully automated sequential immunofluorescence (seqif) for hyperplex spatial proteomics. *Sci Rep*, 13(1):16994, 2023. doi: 10.1038/s41598-023-43435-w.

- Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, and Wan-Yen Lo. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, and Jack Clark. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 2021.
- Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.
- Hunter Lang, David Sontag, and Aravindan Vijayaraghavan. Theoretical analysis of weak-to-strong generalization. *arXiv preprint arXiv:2405.16043*, 2024.
- Nobuyuki Otsu. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975.
- Norberto Malpica, Carlos Ortiz De Solórzano, Juan José Vaquero, Andrés Santos, Isabel Vallcorba, José María García-Sagredo, and Francisco Del Pozo. Applying watershed algorithms to the segmentation of clustered nuclei. *Cytometry: The Journal of the International Society for Analytical Cytology*, 28(4):289–297, 1997.
- Tony F Chan and Luminita A Vese. Active contours without edges. *IEEE Transactions on image processing*, 10(2):266–277, 2001.
- Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2017.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III* 18, 2015.
- Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2019.
- Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical image analysis*, 58:101563, 2019.
- Yuxin Jia, Chao Lu, Xiang Li, Mingxin Ma, Zhongyi Pei, Zhengda Sun, and Yan Chen. Nuclei instance segmentation and classification in histopathological images using a dt-yolact. In *2021 20th International Conference on Ubiquitous Computing and Communications (IUCC/CIT/DSCI/SmartCNS)*, 2021.
- Jacob W Johnson. Adapting mask-rcnn for automatic nucleus segmentation. *arXiv preprint arXiv:1805.00500*, 2018.
- Jacob W Johnson. Automatic nucleus segmentation with mask-rcnn. In *Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 2* 1, 2020.
- Fuxing Long. Microscopy cell nuclei segmentation with enhanced u-net. *BMC bioinformatics*, 21(1):8, 2020.
- Guanbin Lv, Kaigui Wen, Zeyu Wu, Xin Jin, Hong An, and Jun He. Nuclei r-cnn: Improve mask r-cnn for nuclei segmentation. In *2019 IEEE 2nd International Conference on Information Communication and Signal Processing (ICICSP)*, 2019.
- Antti O Vuola, Saad Ullah Akram, and Juho Kannala. Mask-rcnn and u-net ensembled for nuclei segmentation. In *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*, 2019.
- Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support: 4th international workshop, DLMIA 2018, and 8th international workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, proceedings* 4, 2018.
- Lei Ke, Yu-Wing Tai, and Chi-Keung Tang. Deep occlusion-aware instance segmentation with overlapping bilayers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021.
- Yan Kong, Georgi Z Genchev, Xiaolei Wang, Hongyu Zhao, and Hui Lu. Nuclear segmentation in histopathological images using two-stage stacked u-nets with attention mechanism. *Frontiers in Bioengineering and Biotechnology*, 8:573866, 2020.
- Csaba Molnar, Ian H Jermyn, Zoltan Kato, Vilja Rahkama, Päivi Östling, Piia Mikkonen, Vilja Pietiäinen, and Peter Horvath. Accurate morphology preserving segmentation of overlapping cells based on active contours. *Scientific reports*, 6(1):32412, 2016.
- Kallol Roy, Sanjoy Saha, Debapriya Banik, and Debottosh Bhattacharjee. Nuclei-net: A multi-stage fusion model for nuclei segmentation in microscopy images. *Innovations in Systems and Software Engineering*, pages 1–8, 2023.

- Yanning Zhou, Hao Chen, Jiaqi Xu, Qi Dou, and Pheng-Ann Heng. Irnet: Instance relation network for overlapping cervical cell segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I* 22, 2019.
- Hao Jiang, Rushan Zhang, Yanning Zhou, Yumeng Wang, and Hao Chen. Donet: Deep de-overlapping network for cytology instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023.
- Nikita Dvornik, Julien Mairal, and Cordelia Schmid. Modeling visual context is key to augmenting object detection datasets. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
- Debidatta Dwibedi, Ishan Misra, and Martial Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *Proceedings of the IEEE international conference on computer vision*, 2017.
- Hao-Shu Fang, Jianhua Sun, Runzhong Wang, Minghao Gou, Yong-Lu Li, and Cewu Lu. Instaboost: Boosting instance segmentation via probability map guided copy-pasting. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2019.
- Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021.
- Yi Lin, Zeyu Wang, Kwang-Ting Cheng, and Hao Chen. Insmix: towards realistic generative data augmentation for nuclei instance segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2022.
- Xinyi Yu, Guanbin Li, Wei Lou, Siqi Liu, Xiang Wan, Yan Chen, and Haofeng Li. Diffusion-based data augmentation for nuclei image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2023.
- Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2019.
- Hongyi Zhang. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.
- 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: 21st International Conference, Granada, Spain, September 16–20, 2018, Proceedings, Part II* 11, 2018.
- Marius Pachitariu and Carsen Stringer. Cellpose 2.0: how to train your own model. *Nature methods*, 19(12):1634–1641, 2022.
- Carsen Stringer and Marius Pachitariu. Cellpose3: one-click image restoration for improved cellular segmentation. *Nature methods*, pages 1–8, 2025.
- Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: a generalist algorithm for cellular segmentation. *Nature methods*, 18(1):100–106, 2021.
- Michael Y Lee, Jake S Bedia, Salil S Bhate, Graham L Barlow, Darci Phillips, Wendy J Fantl, Garry P Nolan, and Christian M Schürch. Cellseg: a robust, pre-trained nucleus segmentation and pixel quantification software for highly multiplexed fluorescence images. *BMC bioinformatics*, 23(1):46, 2022.
- Can Cui, Ruining Deng, Quan Liu, Tianyuan Yao, Shunxing Bao, Lucas W Remedios, Bennett A Landman, Yucheng Tang, and Yuankai Huo. All-in-sam: from weak annotation to pixel-wise nuclei segmentation with prompt-based finetuning. In *Journal of Physics: Conference Series*, 2024.
- Anwai Archit, Luca Freckmann, Suganya Nair, Nabeel Khalid, Paul Hilt, Vikas Rajashekhar, Marei Freitag, Charlotte Teuber, Gerry Buckley, and Sushmita von Haaren. Segment anything for microscopy. *Nature methods*, pages 1–13, 2025.
- Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature communications*, 15(1):654, 2024.
- Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018.
- Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *International conference on machine learning*, 2019.
- Cheng Tan, Jun Xia, Lirong Wu, and Stan Z Li. Co-learning: Learning from noisy labels with self-supervision. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2021.

- Yongqiang Zhang, Yancheng Bai, Mingli Ding, Yongqiang Li, and Bernard Ghanem. W2f: A weakly-supervised to fully-supervised framework for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.
- Sheng Liu, Kangning Liu, Weicheng Zhu, Yiqiu Shen, and Carlos Fernandez-Granda. Adaptive early-learning correction for segmentation from noisy annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.
- Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.
- Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.
- Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019.
- Neeraj Kumar, Ruchika Verma, Sanuj Sharma, Surabhi Bhargava, Abhishek Vahadane, and Amit Sethi. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE transactions on medical imaging*, 36(7):1550–1560, 2017.
- Yagyensh Chandra Pati, Ramin Rezaifar, and Perinkulam Sambamurthy Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of 27th Asilomar conference on signals, systems and computers*, 1993.