Self-supervised learning (SSL) leverages the underlying data structure to generate supervisory signals for training deep networks. This approach offers a practical solution for learning with multiplexed immunofluorescence brain images where data are often more abundant than human expert annotations. SSL algorithms based on contrastive learning and image reconstruction have demonstrated impressive performances. Unfortunately, these methods were designed and validated mostly on natural images rather than biomedical images. A few recent works have applied SSL to analyzing cell images. However, none of these works studied SSL for multiplexed immunofluorescence brain images. These works also did not provide a clear theoretical justification for adopting a specific SSL method. Motivated by these limitations, our paper presents a self-supervised Dual-Loss Adaptive Masked Autoencoder (DAMA) algorithm developed from the information theory viewpoint. DAMA’s objective function maximizes the mutual information by minimizing the conditional entropy in pixel-level reconstruction and feature-level regression. In addition, DAMA introduces a novel adaptive mask sampling strategy to maximize mutual information and effectively learn brain cell data contextual information. For the first time, we provide extensive comparisons of SSL algorithms on multiplexed immunofluorescence brain images. Our results demonstrate that DAMA is superior to other SSL approaches on cell classification and segmentation tasks. DAMA also achieves competitive accuracies on ImageNet-1k. The source code for DAMA is made publicly available at github.com/hula-ai/DAMA

1. Introduction

Microscopic brain image analysis is critical for medical diagnosis and drug discovery [23]. While collecting large amounts of brain imaging data using high-resolution multiplexed microscopes is efficient, annotating these images is time-consuming and labor-intensive. Each brain slice consists of several hundred thousand cells and dozens of cell types. Labeling these images requires highly skilled biology experts. For these reasons, the number of labels for this application is often limited, while the amount of unannotated data is enormous. Self-supervised learning (SSL) offers a practical solution to this situation.

SSL methods have achieved impressive performance in natural language processing [5, 11, 24, 25], speech processing [3, 26, 32], and computer vision [2, 6, 7, 11, 18]. SSL aims to learn powerful data representations that are useful for the downstream tasks by using pretext tasks created without human supervision. Hence, this approach is ideal for biomedical applications with massive data and limited supervised information. However, there is no best self-supervised method overall [14]; thus, choosing the right SSL learning algorithm is not always straightforward.

A few recent works have applied SSL to cell data [12,27,34]. For instance, [34] reconstructs distorted input to learn representations for quantitative phase image cell segmentation, [12] pre-trains 1M cancer cell images with convolutional autoencoder to classify the drug effects. Miscell [27] utilizes contrastive learning for mining gene information from single-cell transcriptomes. None of these works studies SSL for multiplexed brain cell analysis. In addition, existing works mainly focus on the applications, while the theoretical analysis for adopting a specific SSL method is often unclear. In contrast, this study tries to bridge the gap between biomedical applications and theoretical moti-
We propose a novel SSL framework that optimizes both pixel [4, 17, 36], and feature-level losses [2, 21, 31] for brain cell image analysis, called DAMA, see Fig. 1. Our dual loss is motivated by information theory and the observation that the context around cells is useful for analyzing them correctly. The method maximizes the mutual information between masked inputs [2, 4, 17] and self-supervised signals. Specifically, we first mask the original input and then map it to feature space from where the model learns to reconstruct the original unmasked input. Simultaneously, the exact representations also regress to the representations of a different masked version of the original input. Simultaneously learning pixel reconstruction and feature regression increases the consistency of different masked images from the same input. In addition, from the information standpoint and our observation, we further propose an adaptive masking strategy to enforce the networks to learn better representations.

The main contributions of our paper are as follows:

1. We present a novel SSL method for multiplexed biomedical data analysis, i.e., brain cells, motivated by the information theory perspective. Our method achieved superior performance compared to state-of-the-art SSL methods and supervised.

2. We also propose an adaptive mask sampling strategy that considerably influences learning good representation. This could be the first adaptive masking method for self-supervised learning to the best of our knowledge.

## 2. Related Works

### 2.1. Self-Supervised Learning.

Recently, self-supervised learning (SSL) has exhibited a very successful approach in computer vision [4, 6, 7, 16–18]. However, choosing the right SSL learning algorithm is not always straightforward [14]. For example, one of the characteristics of multiplexed biomedical data is that the context also conveys crucial information about the cell. Learning self-supervised signals as multi-view augmented images with contrastive [7, 16, 18, 21, 31], redundancy reduction [37] or self-distillation [6] objective would discard or unfocus on this context information. As an example, DINO [6] visualizes the attention maps of Vision Transformers (ViT) [13] after training, whose main focus is on the interesting objects and leaves contextual information unattended. MAE [17] and SimMIM [36] learn to reconstruct missing image patches from uncorrupted patches. Similarly, Data2Vec [2] regresses the unmasked patches from masked patches in feature level. However, due to the high random masking ratio, MAE, SimMIM, and Data2Vec would not guarantee to focus on the context information in each iteration. MoCo-v3 [8] learns to increase the mutual information of two augmented views of the same image, e.g., \( I(X_1, X_2) \), due to the effect of augmentation transformations, MoCo-v3 would capture the cell body information and abandon the surrounding context information which distinct for each cell. Alternatively, based on ViT [13] framework, we optimize the objective function on both pixel-level reconstruction [4, 17, 36] and features-level regression [2] to predict the content of masked regions. By doing so, the algorithm will concentrate on invariant features and the entire image.

### 2.2. Masked Image Modeling (MIM).

Recent works built upon Vision Transformer (ViT) [13] framework, such as BeiT [4], MAE [17], SimMIM [36] have shown potential of MIM in learning representations. Similar to our work, these prior studies propose masking out a random subset of image patches and encourage reconstructing the original pixel, but our work differs in that we also introduce regress feature representations of multiple ViT blocks [2]. On the other hand, our method is also distinct from Data2Vec [2] as they take the masked and unmasked patches as input and predict features produced from uncorrupted input. We, however, apply only to the visible patches and predict the feature also produced from the visible patches of the second network, i.e., teacher or momentum network. Another point of separating our work from others is that we introduce an adaptive masking strategy that can learn better representation and boost fine-tune performance.

### 2.3. Self-Supervised Learning on Biomedical Data

Available SSL methods are usually applied to specific applications with less novelty contribution in the biomedical field. For instance, [34] reconstructs distorted input to better representations of quantitative phase image cell segmentation, [12] pre-trains 1M cancer cell images with convolutional autoencoder to classify the drug effects. Misee [27] utilizes contrastive learning for mining gene information from single-cell transcriptomes. In addition, there are very few papers that study multiplexed biomedical data. While application-centric studies are acceptable, the theoretical analysis for adopting a specific SSL method is often unclear, missing out on potential approaches. In contrast, this study aims to bridge the gap between biomedical applications and theoretical motivation; and apply it to multiplexed biomedical data, e.g., brain cell data.

### 3. Self-Supervised Learning from Information Theory Perspective

**Notations.** For the rest of this paper, we denote the input and self-supervised signal in general as \( X \) and \( S \), respectively. \( S \) can be the augmented image [6, 7, 18] or the...
target of image reconstruction [4,17,36]. The deterministic mapping function $F$ maps the input $X$ to its representations $Z_X$, i.e., $Z_X = F(X)$, and function $G$ reconstructs the input as $S = G(Z_X)$. Regarding the information, we use $I(A, B)$, $H(A)$, and $H(A|B)$ to denote the mutual information, entropy, and conditional entropy of variables $A$ and $B$, respectively.

As shown in Fig. 1 (b), solid and dotted rectangles represent the information of input $X$ and self-supervised signal $S$, respectively. From the information theory perspective, the mutual information between the representation $Z_X$ and $S$, denoted as $I(Z_X, S)$ (grey area), measures the amount of information obtained about one from the knowledge of the other. This mutual information can be expressed as the difference between two entropy terms:

$$I(Z_X, S) = H(Z_X) - H(Z_X|S) = H(S) - H(S|Z_X) \quad (1)$$

In the self-supervised learning context, one can directly maximize $I(Z_X, S)$ like in Compler [21]. Alternatively, minimizing the conditional entropy $H(S|Z_X)$ (green area) [18,32] would also encourage $S$ to be fully determined by $X$, indirectly maximizing $I(Z_X, S)$ and minimizing the irrelevant information between $X$ and $S$. The Eq. (1) can be interpreted as $I(Z_X, S)$ minimize the uncommon information between $X$ and $S$ [31,37]. Hence, if $X$ and $S$ are independent, then $I(Z_X, S) = 0$, while if $X$ and $S$ are related, then $I(Z_X, S)$ will be greater than some lower bound. For this reason, $S$ is usually the augmented images [2,6,7,16,18,31,37] or random masked images [2,4,17,36].

Augmentation could boost the performance of self-supervised learning algorithm [7,16,18]. From the Information Bottleneck principle [28,29], augmented images could enforce the encoder $F$ to estimate invariant information [35]. However, augmentation is data-dependent, and finding the right transformation could sometimes be inconvenient. SimCLR [7] conducts a resource-consuming experiment with the combination of only two transformations to find the most favorable combination for ImageNet-1k [10]. Moreover, augmentation could remove contextual information, which is important for downstream biomedical tasks.

Motivated by [21,31], our method aims to minimize the conditional entropy $H(S|Z_X)$. Fig. 1 (a) provides an illustration of our method. While [31] employs forward-inverse predictive learning to boost the performance of contrastive objective, [21] benefits from dual prediction and contrastive learning for recovering missing views. We, however, take a fundamentally different approach by not targeting to optimize the contrastive function [21,31] but focusing on maximizing the mutual information between masked inputs and self-supervised signals at pixel-level reconstruction [4,17,36] and features-level regression [2]. In addition, while most existing works [2,4,17,36] utilize a random image masking strategy, our method uses adaptive sampling to more effectively minimize the conditional entropy $H(S|Z_X)$ and learn better representations. To the best of our knowledge, our method is the first to use adaptive image masking for self-supervised learning.
4. Dual-loss Adaptive Masked Autoencoder

Motivated from the information theory, this section proposes a Dual-loss Adaptive Masked Autoencoder (DAMA) for self-supervised learning. Our method optimizes an objective function associated with pixel- and feature-level information masking. As illustrated in Fig. 1 (a), it consists of a dual objective function:

$$L_{\text{total}} = L_p + \alpha L_f$$  \hspace{1cm} (2)

where $L_p$ and $L_f$ are the losses associated with pixel-level reconstruction and feature-level regression, respectively. $\alpha$ is a non-negative constant. From the information theory perspective, our method optimizes $I(Z_X, S)$ and $H(S|Z_X)$. In addition, we present a novel adaptive masking strategy that is better than random masking in terms of performance and theory background. We first provide context information related to the method development and introduce theoretical details later. In our implementation, we fixed $\alpha = 1$ for all experiments.

DAMA uses a Vision Transformer (ViT) [13] as the backbone network. Given a multiplexed input image with seven channels $x \in \mathbb{R}^{H \times W \times 7}$, we reshape it into small patches $(x^n)_{n=1}^N$, where $N = HW/P^2$ patches and $P$ is the resolution of each patch. We masked $m\%$ of the patches and denote them as $\mathcal{M} = \{1, \ldots, N\}^{m \times N}$. Here, unlike BEiT [4] and Data2Vec [2] treat the masked patches and unmasked patches as input to ViT, i.e. $x^P = \{x_i^P : i \notin \mathcal{M}\}_{i=1}^N \cup \{e_i^P : i \in \mathcal{M}\}_{i=1}^N$, where $e_i^P$ is the learnable embedding replacing for masked patches, we feed only the unmasked patches $x^U_i = \{x_i^P : i \notin \mathcal{M}\}_{i=1}^N$ which similar to MAE [17].

4.1. Pixel-level Reconstruction

Here, we present the theoretical background for the pixel-level loss $L_p$ in Eq. (2). In the context of pixel reconstruction, we regard the self-supervised signal $S$ as the reconstruction target, i.e., the original input, denoted as $S_X$ in Fig. 1(a).

According to Eq. (1), minimizing the conditional entropy $H(S|Z_X)$ (green area) would also encourage $S$ to be fully determined by $X$, indirectly maximize $I(Z_X, S)$, and minimize the irrelevant information between $X$ and $S$ [31, 37]. To do so, the learned representation $Z_X$ is encouraged to reconstruct the self-supervised signal $S$, which leads to maximizing the log conditional likelihood: $-H(S|Z_X) = \mathbb{E}_{P_S, Z_X} [\log P(S|Z_X)]$. However, directly inferring $P(S|Z_X) = \frac{P(Z_X|S)P(S)}{P(Z_X)}$ would be intractable.

A common approach to approximate this objective is to define a variational distribution $Q(S|Z_X)$ and maximize the lower bound $\mathbb{E}_{P_S, Z_X} [\log Q(S|Z_X)]$ using variational information maximization technique [1] as:

$$I(Z_X, S) = H(S) - H(S|Z_X) = \mathbb{E}_{P_S, Z_X} [\log P(S|Z_X)] + H(S)$$

$$= D_{KL} (P(S|Z_X) \parallel Q(S|Z_X)) \geq 0$$

$$+ \mathbb{E}_{P_S, Z_X} [\log Q(S|Z_X)] + H(S) \geq \mathbb{E}_{P_S, Z_X} [\log Q(S|Z_X)]$$

(3)

Such $Q(\cdot)$ can be any type of Gaussian [15, 21, 31], Laplacian [38], categorical [9] distribution, or neural network [22, 33]. We present the $Q(S|Z_X)$ as Gaussian distribution with $\sigma I$ as diagonal matrix, i.e., $\mathcal{N}(S|G(Z_X), \sigma I)$ [15, 21, 31], where $G(\cdot)$ is parameterized as deterministic mapping model which map $Z_X$ to the sample space of $S$. After specifying the mapping functions $G$, the maximizing $\mathbb{E}_{P_S, Z_X} [\log Q(S|Z_X)]$ objective functions in pixel-level is:

$$L_p = \min \sum_{i=1}^\alpha [\mathbb{E}_{P_S, Z_X} \|G(Z_X^m) - S\|_2^2], \ i = 1, 2$$

(4)

where $Z_X^m$ is the representation of masked input $X_i^m$, i.e. $Z_X^m = F(X_i^m)$, and $i = 1, 2$ represents for two branches of models as shown in Fig. 1. From the above objective function, we can notice that when $L_p = 0$, i.e., $S$ can be fully determined by $Z_X^m$, mathematically, $H(S|Z_X)^m = 0$. One common situation is that the $G(F(\cdot))$ becomes the identical mapping $I$, and the network will learn nothing. For this reason, $X^m$ is usually the augmented images or random masking images from the same source to avoid the degenerate solution. Our DAMA employs the image masked autoencoder modeling approach similar to [2, 4, 17, 36] instead of augmenting the input. To leverage the masking operation to contribute more than just generating random masks, we propose a novel adaptive masking strategy that can increase the mutual information $I(Z_X, S)$ and learn better representation, whose details will be explained in section 4.3.

4.2. Feature-level Regression

To further encourage maximizing the mutual information $I(Z_X, S)$, DAMA also consist a feature-level regression objective $L_f$ in Eq. (2). In the context of feature-level regression, we prefer self-supervised signal $S$ as the feature target produced by $F_Z(X^m_2)$ and indicate as $S_2$ in Fig. 1. DAMA predicts feature representations of masked view $X^m_2$ based on masked view $X^m_1$, i.e., $F_Z : X^m_1 \rightarrow X^m_2$, where $F_Z$ is the mapping function. This is different from Data2Vec [2] which predicts feature representations of the original uncorrupted input $X$ based on a masked view $X^m$ in a student-teacher setting, i.e., $F_Z : X^m \rightarrow X$. Furthermore, the masked patches of $X^m_2$ are decided by adaptive sampling strategy.

Similar to the pixel-level loss, maximizing the mutual information $I(Z_X, S)$ leads to maximizing the log conditional likelihood $\mathbb{E}_{P_S, Z_X} [\log P(S|Z_X)]$. We can in-
introduce a variational distribution \(Q(S|Z_{X})\) and maximize the lower bound \(\mathbb{E}_{P_{S, Z_{X}}} \left[ \log Q(S|Z_{X}) \right]\). Let \(Q(S|Z_{X})\) as Gaussian distribution with \(\sigma I\) as diagonal matrix, i.e., 
\(\mathcal{N}(S|F_{Z}(Z_{X}), \sigma I)\). The objective function is obtained as:

\[
\mathcal{L}_{f} = \min \mathbb{E}_{P_{S, Z_{X}}} \left\| F_{Z}(Z_{X}^{m}) - S \right\|_{2}^{2}
\]

where \(Z_{X}^{m}\) is the representation of masked input \(X_{1}^{m}\), i.e., \(Z_{X}^{m} = F(X_{1}^{m})\). Note that \(F\) and \(F_{Z}\) are two different mapping functions; refer to Fig. 1 for visual illustration. Given that our application of interest is brain image analysis, where context provides critical information for classification and segmentation tasks, we adopt the smooth L1 loss as Gaussian distribution with \(\sigma I\) as diagonal matrix, i.e., \(\mathcal{N}(S|F_{Z}(Z_{X}), \sigma I)\). Hence, the Eq. (5) becomes:

\[
\mathcal{L}_{f} = \begin{cases} 
\frac{1}{2}(F(Z_{X}^{m}) - S)^{2} / \beta, & |F(Z_{X}^{m}) - S| \leq \beta \\
\frac{1}{2}F(Z_{X}^{m}) - S, & \text{otherwise}
\end{cases}
\]

where \(\beta\) is the smoothing from L2 to L1 loss term and depends on the difference between \(F_{Z}(Z_{X}^{m})\) and \(S\). In addition, the self-supervised signal \(S\) is taken from the last \(K\) blocks of the second branch of the model before normalization to each block and then averaging similarly as in Data2Vec. In our implementation, \(K = 6\) and \(\beta = 2\) for all experiments.

4.3. Adaptive Masking Strategy

Unlike other works [2, 4, 17, 36], we propose an adaptive masking strategy to produce masked images \(X_{2}^{m}\). This strategy helps increase the mutual information \(I(Z_{X}, S)\) and learn better representations. The method is originated from our observation of the theoretical background presented in the pixel-level reconstruction section 4.1. The patches with the highest loss indicate the lowest mutual information \(I(Z_{X}^{m}, S)\). See Algorithm 1.

The proposed strategy takes the random binary mask of \(X_{1}^{m}\) and the patch reconstruction loss in Eq. (4) as inputs. It selects the patches with the highest loss, which indicate the lowest mutual information \(I(Z_{X}^{m}, S)\) as masked patches for \(X_{2}^{m}\). Regarding the unmasked patches in \(X_{1}^{m}\), based on the overlap ratio, some will become the unmasked patches, and the rest will serve as masked patches in \(X_{2}^{m}\). The overlap ratio is fixed at 50% for all experiments. This guarantees that the feature-pixel regression would not be too difficult to predict. One can think of the adaptive image masking strategy as a collaboration between two students, where the first student estimates the difficulties of reconstructing different patches, and the second student uses that information to select challenging patches to enhance the performance. Note that we develop DAMA upon ViT framework [13]. Hence, we compute the reconstruction loss patch-wise, and unmasked patches are not considered in computing loss [2, 4, 17, 36].

Algorithm 1 Pytorch-like Adaptive Masking Pseudocode

```python
def adaptive_mask(m1, loss, mask_ratio, overlap_ratio):
    # mask_ratio: masking ratio in [0, 1]
    # m1, m2: binary masks; size[N, L]
    # overlap_ratio: overlap ratio between 2 masks
    # loss: patch reconstruction losses; size[N, L]
    # N: batch size
    # L: total number of patches in images
    len_keep = int(L * (1 - mask_ratio))
    loss_len = int(L - len_keep * 2)
    overlap_len = int(len_keep * overlap_ratio)
    # get ids of high loss patches
    # discard losses of unmasked patches in m1
    loss_sorted = argsort(loss)
    loss_ids=loss_sorted[:,-(loss_len+overlap_len):]
    # m1(1) becomes m2(0) and m1(0) becomes m2(1)
    m2 = where(m1 == 1, 0, 1)
    # assign ids of highest loss m1 to m2 as masks
    m2[where(loss_sorted[:, -((loss_len+overlap_len)):])] = 1
    # overlap of unmasked patches of m1 and m2
    m2 = m2[[i, j](m1_ids), (i, j)(m2_ids)]
    return m2
```

5. Experiments

In this section, we validate our DAMA algorithm on the multiplexed immunofluorescence brain image dataset and compare its performance to supervised learning and state-of-the-art SSL approaches. See Table 1, Table 2, Table 3, and Fig. 2. See Appendix for extensive experiments and results in Table 5, Fig. 6, 7, 8, 9, and 10.

5.1. Experimental Settings

5.1.1 Brain cell dataset.

The cell images are collected from 5 major cell types from rat brain tissue sections: neurons, astrocytes, oligodendrocytes, microglia, and endothelial. Seven biomarkers are applied as the feature channels: DAPI, Histones, NeuN, S100, Olig2, Iba1, and RECA1. DAPI and Histones are utilized to reveal the cells’ locations, while other biomarkers are useful for identifying cell types. All models take 7-channel images as the input. No cell detection pre-processing has been applied to these images; thus, some patches may contain more than one cell, but only the central cell is what we are interested in, and all other cells are generally considered as the background for the cell type classification task, see the Fig. 11 for example cell images. Since cell types are highly corresponded with specific biomarker, we only use affine transformations, such as rotating, translating, flipping, scaling, and no colorizing transformation.

For pretraining and evaluating the SSL methods on the classification task, we collected 1600 cells images for each cell type. In a total of 8000 images, our biologists manu-
Table 1. Comparisons of finetuning results with state-of-the-art SSL methods and randomly initialized in accuracy and error rate. Training epochs and training times are listed along with the methods.

| Methods            | Box mAP | Box mAP@50 | Box mAP@75 | Mask mAP | Mask mAP@50 | Mask mAP@75 |
|--------------------|---------|------------|------------|----------|-------------|-------------|
| ViT random init.   | 63.4    | 90.8       | 73.9       | 66.7     | 90.9        | 76.1        |
| Swin random init   | 63.2    | 90.6       | 73.7       | 66.3     | 90.5        | 76          |
| MAE 800            | 63.7    | 90.8       | 74.8       | 67.1     | 91.4        | 76.9        |
| MAE 1600           | 63.8    | 90         | 73.3       | 66.3     | 90.1        | 76.3        |
| MOCO-v3 500        | 63.2    | 90.5       | 73.2       | 66.5     | 91          | 75.9        |
| MOCO-v3 1000       | 63.1    | 90.2       | 73         | 66.1     | 90.8        | 75.2        |
| SIMMIM-ViT 800     | 63.6    | 91.1       | 74.1       | 66.9     | 91.1        | 76.1        |
| SIMMIM-Swin 800    | 64.2    | 91.3       | 75.1       | 67       | 91.2        | 77          |
| DAMA 500           | 64.1    | 91.1       | 74.2       | 67.2     | 91.1        | 77          |
| DAMA 1000          | **64.6**| **91.4**   | **75.3**   | **67.3** | **91.3**    | **77**      |

Table 2. Comparisons of segmentation accuracy with state-of-the-arts on brain cell dataset. Bold and underlined are the highest and second highest scores for each column.

5.2. Implementation Details

We implemented DAMA using Pytorch. Unless stated otherwise, we trained on ViT-Base used Adam optimizer [20] with base learning rate of 0.00015 [8], batch size of 512, image size 128×128×7, ViT patch size 16. Regarding state-of-the-arts implementation, we take the official released code [8, 17, 30] and conduct pre-training with our biomedical data, except for Data2Vec [2]. We also use ViT-Base framework and similar parameters as above for these experiments. We report results of our DAMA and MAE [17] with masking ratios 80% and 60% for Data2Vec [2]. Training epochs and training times are listed along with the methods in result tables. All experiments were done on 4 GPUs of V100 32GB. Pre-training or finetune experiments of different methods on the same dataset have the same random seed.

5.3. Comparisons with State-of-the-arts

5.3.1 Brain cell dataset.

In Table 1, we compare the finetuning classification accuracy/error rate of randomly initialized and pretrained self-supervised models. Our DAMA outperforms both state-of-the-art SSL and randomly initialized methods.

For the segmentation task, our biologists have manually labeled 4000 images for model finetuning, and we keep the remaining 4000 images for pretraining purposes. For each finetune experiment, we randomly shuffle images and then split the finetune set into 60%/40%, i.e., 2400/1600, for training/validation sets. We validate each SSL method by repeating the finetune experiment ten (10) times and averaging the results.

For the segmentation task, our biologists have manually collected and annotated 181 images of size 512×512×7. We split these images into the size of 128×128×7, totaling 2896 images. We further split these images into training/val sets for evaluating the segmentation task with a ratio of 60/40. We augmented the training set for pretraining SSL methods. The segmentation task requires segmenting the cell body from the background regardless of the cell type.
MoCo-v3 (95.47% versus 91.75%). Moreover, initializing the model with MoCo-v3 representation does not improve the accuracy (91.75%) over a randomly initialized classifier (91.98%). This suggests MoCo-v3 introduces biases on small datasets. In addition, DAMA’s pre-training times (5h) are not much higher than MAE’s (4h) as DAMA uses another network to learn the second mask. Data2Vec (4h) only leverages features for learning, leading to less training time. MoCo-v3’s training times (6h) are slower than MAE, Data2Vec, and our DAMA adaptive masking method.

Cell Segmentation. Pretrained and finetuned with the same ViT backbone architecture, except for SimMIM-Swin, and object detector frameworks Mask R-CNN [19], our DAMA achieves the best performances compared to other SSL methods, see Table 2 and Fig. 2. Cell segmentation is challenging as no single biomarker determines the whole cell body. DAMA significantly outperforms all baseline methods. We hypothesize that DAMA is more capable of utilizing contextual information around cells to resolve ambiguous cases where cells are dense and overlapping. Two observations can support this hypothesis. First, MoCo-v3’s objective is to increase the mutual information of two augmented views coming from the same image, e.g., \( I(X_{\text{mask}}, X_{\text{original}}) \). The random augmentation operations encourage MoCo-v3 to focus on the cell body while discarding the contextual features since they are likely to decrease the mutual information. Second, MAE and SimMIM learn to reconstruct missing image patches from un-corrupted patches, e.g., \( I(X_{\text{mask}}, X_{\text{original}}) \). However, the high random masking ratios in these two methods (e.g., 0.8) substantially reduce the contextual information. In contrast, our DAMA model emphasizes contextual information by adaptive masking strategy in each iteration and achieves the best segmentation performance. Fig. 3 shows that DAMA can segment clusters of cells better than other methods. The results demonstrate DAMA’s significant advantage in analyzing multiplexed brain cell data. More visualization results are in Fig. 6.

Fig. 2 shows the segmentation performances of DAMA and other methods. For DAMA, overall AP at IoU@.75 is 0.77, and perfect localization \( Loc \) (PR at IoU@.1) increases AP to 0.956. Since we only segment the cell body from the background regardless of the cell type, e.g., having only one class, the super categories \( Sim \) and classes confusion \( Oth \) are unchanged compared to perfect localization \( Loc \). Removing background false positives \( BG \) would increase the performance by 0.014 (to 0.97 AP), and the rest of the errors are missed detections. In summary, DAMA’s errors come from imperfect localization \( Loc \) and confusing background \( BG \). Regarding other methods, the amount of \( Loc \) and \( BG \) errors are higher than those of DAMA. The meaning of errors is detailed on the COCO website cocodataset.org/#detection-eval. We also present the detection and segmentation errors in Fig. 7 and 8, the precision-recall curves at different IoU threshold in Fig. 9, 10 in Appendix section.
Methods | Pre-trained epochs | Acc.  
---|---|---  
Moco-v3 | 600 | 83.2  
BEiT | 800 | 83.4  
SimMIM | 800 | 83.8  
Data2Vec | 800 | 84.2  
DINO | 1600 | 83.6  
iBOT | 1600 | 84.0  
MAE | 1600 | 83.6  
Our DAMA | 500 | 83.17  

Table 3. Comparisons results of DAMA and state-of-the-arts on ImageNet-1k. DAMA is pretrained and fine-tuned with image size of $224 \times 224$.

| Masking strategy | Model settings | Masking ratio %  
---|---|---  
Rand. w overlap | student-ema | 60 | 70 | 80 | 90  
Adapt. w/o overlap | student-ema | 94.65 | 94.67 | 94.76 | 94.65  
Adapt. w overlap | student-ema | 95.14 | 95.4 | 95.43 | 95.06  
| shared weights | 94.81 | 95.13 | 95.13 | 94.65  
| student1-student2 | 95.34 | 94.46 | 94.47 | 95.38  

Table 4. Effect of sampling strategies, model settings, and masking ratios in term of classification accuracy. **Bold** are the highest score for each column, while **underlined** are highest score for each row.

Figure 4. Pre-training curves of different model settings on adaptive masking condition: Pixel-level reconstruction loss in (a) network1 and (b) network2, (c) feature-level regression loss, and (d) total loss.

5.3.2 ImageNet-1k.

To demonstrate the potential of DAMA on other natural image type, we present DAMA’s result on ImageNet-1k [10], see Table 3. DAMA is competitive to other state-of-the-art algorithms despite smaller numbers of pre-trained epochs and without any ablation experiment for searching optimal hyper-parameters. Due to the computational resource needed for training on such a large-scale dataset, we perform only a single pretraining/finetuning experiment on ImageNet-1K with the same configuration as for training on the brain image dataset, except for the image size and pretraining/finetuning batch size as $224 \times 224 \times 3$, and $4096/1024$, respectively.

5.4. Ablation Studies

In Fig. 4, Fig. 5, and Table 4, we ablate our DAMA with different masking strategies, model strategies, and masking ratios. See Appendix for more results.

**Masking Strategy and Masking Ratio.** We compare our DAMA with *adaptive* and *random* masking strategies and analyze how they affect the finetuning results. We vary the masking ratio in the range of (60%-90%). The masking ratio of 80% achieves a better result for all three masking settings. Similarly, MAE reports high masking ratios, ideally ranging in $60\%-80\%$ for good finetuning performance on ImageNet-1K. Primarily, adaptive masking produces better accuracy with different masking ratios in the same training condition. These experiments justify the effectiveness of our method. Overlapping, a step in the proposed adaptive masking, means some unmasked patches in $X_{m1}$ will also be unmasked in $X_{m2}$. The overlapping ratio is set at 50% for all experiments. No overlapping between unmasked input $X_{m1}$ and $X_{m2}$ leads to more challenging optimization of $L_f$ as illustrated in Fig. 4. Specifically, the loss curves for none overlapping experiments are higher than others. These results demonstrate the effect of adaptive masking and feature-level regression on the overall objective function.

**Student and Teacher Configurations.** We perform ablation studies with different variants of student and teacher
network configurations. Specifically, we compare three configurations: student-teacher, shared-weights (student1-student1), and student1-student2. student-teacher network is an exponential moving average (EMA) on the student weights [18], and the update rule is \( \theta_t \leftarrow \lambda \theta_t + (1 - \lambda) \theta_s \), where \( \lambda \) follows cosine scheduled from 0.996 to 1 during training [6, 16, 18]. Both networks use the same set of parameters in the shared-weights setting. The student1-student2 setting consists of two independent networks, allowing more “freedom” in optimizing objective functions. Table 4 and Fig. 4 show that the student-teacher is less effective than shared-weights (student1-student1) and student1-student2 settings. In addition, utilizing adaptive sampling strategy can improve the performance with the same network configuration. This supports the advantage of our method.

Choice of \( \alpha \). The choice of parameter \( \alpha \) usually is case-by-case determined. Following [21, 31], we keep the scale of feature-level reconstruction loss \( L_f \) to \( \frac{1}{10} \) to \( \frac{1}{100} \) to the scale of the pixel-level reconstruction loss \( L_p \), see Fig. 4. We empirically found that setting \( \alpha = 1 \) yields consistently good performance. The paper [31] shares a similar approach to our paper, except it has an additional contrastive loss, e.g., \( \mathcal{L}_{total} = \alpha_1 \mathcal{L}_p + \alpha_2 \mathcal{L}_f + \alpha_3 \mathcal{L}_{CL} \). They construct controlled experiments to study the effect of this sensitive and dataset-dependent hyper-parameters.

Effect of Adaptive Masking Strategy on Learning Contextual Information. We show in Fig. 5 the reconstruction results of (b) MAE and (c-g) our five settings on the same setting as in Table 4. Except for (a), in (b-g), the upper row is the same random masking input \( X_{m}^n \) (left) and its reconstruction result \( G(Z_{X_{m}^n}) \) (right), the lower row is the applied mask strategy input \( X_{m}^n \) (left) and its reconstruction result \( G(Z_{X_{m}^n}) \) (right). From left to right, (a) original image is center cropped as input green box; (b) MAE’s result, (c) our model-ema random sampling result; our adaptive sampling: (d) student-teacher without mask overlapping; (e) student-teacher with mask overlapping; (f) share-weights with mask overlapping; and (g) student1-student2 with mask overlapping. Without adaptive masking, (c) can not reconstruct properly. This could be explained as the teacher network is constrained by its student network and could not produce reasonable inference with another random mask. In contrast, (d) also uses the student-teacher setting, but with the adaptive mask, can reconstruct the cell reasonably well. Note that, while MAE can not reconstruct the region in red box which is part of another cell in a single iteration, our method can do that with adaptive sampling. This suggests MAE could leave out these fine-grain details even with several epochs in a random high masking ratio setting, i.e., 80%. Conversely, our DAMA combines pixel- and feature-level optimization with adaptive masking to identify those details in every iteration with a high masking ratio. This supports the advantage of our method.

6. Conclusion

This paper introduces DAMA as the first adaptive masking SSL for learning effective representations from multiplexed immunofluorescence brain images. DAMA leverages a dual loss consisting of a pixel-level reconstruction and a feature-level regression. Our experiments show DAMA’s competitive performance on both cell classification and segmentation tasks. Our work demonstrate the importance of adaptive mask sampling and information-theoretic dual loss function in SSL. DAMA also achieves competitive accuracy on ImageNet-1k dataset with a single experiment.

References

[1] David Barber Felix Agakov. The im algorithm: A variational approach to information maximization. Advances in neural information processing systems, 16(320):201, 2004. 4
[2] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. arXiv preprint arXiv:2202.03555, 2022. 1, 2, 3, 4, 5, 6
[3] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems, 33:12449–12460, 2020. 1
[4] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254, 2021. 2, 3, 4, 5
[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020. 1
[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 9650–9660, 2021. 1, 2, 3, 9
[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In International conference on machine learning, pages 1597–1607. PMLR, 2020. 1, 2, 3
[8] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 9640–9649, 2021. 2, 6
[9] Zihang Dai, Zhihui Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. arXiv preprint arXiv:1901.02860, 2019. 4
[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image
[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[12] Andrei Dmitrienko, Mauro Miguel Masiero, and Nicola Zamboni. Self-supervised learning for analysis of temporal and morphological drug effects in cancer cell imaging data. 2021.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

[14] Linus Ericsson, Henry Gouk, and Timothy M Hospedales. How well do self-supervised models transfer? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5414–5423, 2021.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. Advances in neural information processing systems, 27, 2014.

[16] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Ghaslaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. Advances in Neural Information Processing Systems, 33:21271–21284, 2020.

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Pieter Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. arXiv:2111.06377, 2021.

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 9729–9738, 2020.

[19] Kaiming He, Georgia Gkioxari, Pieter Dollár, and Ross Girshick. Mask r-cnn. In Proceedings of the IEEE international conference on computer vision, pages 2961–2969, 2017.

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

[21] Yijie Lin, Yuanbiao Gou, Zitao Liu, Boyun Li, Jianming Liu, and Xi Peng. Completer: Incomplete multi-view clustering via contrastive prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11174–11183, 2021.

[22] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3431–3440, 2015.

[23] Dragan Maric, Jahanadur Jahanipur, Xiaoyang Rebecca Li, Aditi Singh, Aryan Mobiny, Hien Van Nguyen, Andrea Sedlock, Kedar Grama, and Badrinath Roysam. Whole-brain tissue mapping toolkit using large-scale highly multiplexed immunofluorescence imaging and deep neural networks. Nature communications, 12(1):1–12, 2021.

[24] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018.

[25] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9, 2019.

[26] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. arXiv preprint arXiv:1904.05862, 2019.

[27] Hongru Shen, Yang Li, Mengyao Feng, Xilin Shen, Dan Wu, Chao Zhang, Yichen Yang, Meng Yang, Jian Hu, Jilei Liu, et al. Miscell: An efficient self-supervised learning approach for dissecting single-cell transcriptome. Iscience, 24(11):103200, 2021.

[28] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. arXiv preprint physics/0004057, 2000.

[29] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In 2015 ieee information theory workshop (itw), pages 1–5. IEEE, 2015.

[30] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In International Conference on Machine Learning, pages 10347–10357. PMLR, 2021.

[31] Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Self-supervised learning from a multi-view perspective. International Conference on Learning Representations, 2020.

[32] Aaron Van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. arXiv e-prints, pages arXiv–1807, 2018.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

[34] Tomas Vicar, Jiri Chmelik, Roman Jakubicek, Larisa Chmelikova, Jaromir Gumulec, Jan Balvan, Ivo Provaznik, and Radim Kolar. Self-supervised pretraining for transferable quantitative phase imaging cell segmentation. Biomedical optics express, 12(10):6514–6528, 2021.

[35] Mike Wu, Chengxu Zhuang, Chao Zhang, Yichen Yang, Mengyang Qian, Jiani Hu, Jilei Liu, et al. Miscell: An efficient self-supervised learning approach for dissecting single-cell transcriptome. Iscience, 24(11):103200, 2021.

[36] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In 2015 ieee information theory workshop (itw), pages 1–5. IEEE, 2015.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

[38] Tomas Vicar, Jiri Chmelik, Roman Jakubicek, Larisa Chmelikova, Jaromir Gumulec, Jan Balvan, Ivo Provaznik, and Radim Kolar. Self-supervised pretraining for transferable quantitative phase image cell segmentation. Biomedical optics express, 12(10):6514–6528, 2021.

[39] Mike Wu, Chengxu Zhuang, Chao Zhang, Yichen Yang, Mengyang Qian, Jiani Hu, Jilei Liu, et al. Miscell: An efficient self-supervised learning approach for dissecting single-cell transcriptome. Iscience, 24(11):103200, 2021.
A. Appendix

A.1. Noisy datasets

**Real-30k and Real-170k Brain Cell Dataset.**

Augmentation could generate unlimited data. However, the underline structure of data is likely to remain the same. To exam our method on noisy data, we first cropped the large image into many $1000 \times 1000$ images and performed morphological transformations, i.e., erosion. These images were then applied watershed segmentation to identify the cells’ location. From cells’ center locations, we cropped with the size of $100 \times 100$ to get the images. To be compatible with the manually collected set, we collected 30k random cell images regardless of the cell type as the second training set, called Noisy-30k brain cells dataset. In addition, we further constructed another Noisy-170k brain cells dataset as a large-scale dataset.

Our DAMA achieves competitive results better than MAE and Data2Vec and is more stable on three brain cell datasets. MoCo-v3 is influenced by a small pretraining dataset but improves on a large dataset. Table 5 presents the comparisons of DAMA and other methods on Noisy-30k and Noisy-170k. MoCo-v3 has the best performance compared to others. This suggests that a small pretraining set has a negative impact on MoCo-v3. The results also indicate larger pretraining data and longer training time. MoCo-v3 can outperform other methods. This is expected since MoCo-v3 learns to increase the mutual information of two augmented views of the same image, e.g., $I(X_1, X_2)$, capturing better cell body information that is invariant across the dataset. In addition, the classification is considered not a rigorous task by utilizing the one-to-one cidence between cell types and biomarkers. However, this is also a downside since MoCo-v3 would abandon other critical information, e.g., contextual information. Data2Vec does not produce good results since it only learns to regress from low dimension features. On the other hand, MAE and DAMA have better results and are comparable to those from the original dataset.

A.2. Precision-Recall Curves

We present the overall-all-all precision recall curves for bounding box and segment mask in Fig. 7 and 8. DAMA’s results are come from imperfect localization Loc and background confusions BG. Regarding other methods, the amount of Loc and BG errors are higher than those of DAMA. Note that, we has only one class, i.e., segment cell body from background regardless its type.

The precision recall curves at different IoU threshold for bounding box and segment mask are shown in Fig. 9 and 10. For both detection and segmentation, DAMA has the best scores at the IoU from $0.1 : 0.75$ and are competitive at $0.8 : 0.9$. 

[38] 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.
| Folds   | 0  | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | Avg. ↑ | Err. ↓ |
|---------|----|----|----|----|----|----|----|----|----|----|--------|--------|
| Random init. | 91.75 | 91.19 | 92.75 | 92.69 | 92.56 | 92.31 | 91.44 | 91.06 | 93  | 91.06 | 91.98(+0.00) | 8.02   |
| Data2vec 800 (4h) | 93.69 | 93 | 93.31 | 94.06 | 93.56 | 94.19 | 93.38 | 93.31 | 93.81 | 93.5 | 93.58(+1.60) | 6.42   |
| Data2vec 1600 (8h) | 91.5 | 90.69 | 92.81 | 92.38 | 92.62 | 92.62 | 91.56 | 91.94 | 92.19 | 91.5 | 91.98(+0.00) | 8.02   |
| MOCO-v3 500 (6h) | 95 | 94.12 | 95.81 | 96 | 95.75 | 95.19 | 95.12 | 94.44 | 95.5 | 95.19 | 95.21(+3.23) | 4.79   |
| MOCO-v3 1000 (12h) | 94.44 | 93.62 | 94.38 | 95.19 | 94.69 | 95.25 | 94.12 | 94.38 | 95.25 | 94.31 | 94.56(+2.58) | 5.44   |
| MAE 800 (4h) | 94.81 | 94.44 | 94.56 | 94.81 | 94 | 94 | 94.38 | 94 | 94.88 | 93.69 | 94.35(+2.37) | 5.65   |
| MAE 1600 (8h) | 94.38 | 94.19 | 95.12 | 95.19 | 94.44 | 93.94 | 94.12 | 94.19 | 94.44 | 93.94 | 94.49(+2.51) | 5.51   |
| DAMA 500 (5h) | 95 | 94.44 | 95.75 | 95.69 | 94.69 | 95.19 | 94.38 | 95.25 | 94.81 | 95.07(+3.09) | 4.93   |
| DAMA 1000 (10h) | 94.5 | 94.06 | 95.75 | 95.44 | 94.69 | 94.44 | 94.38 | 95.25 | 94.25 | 94.72(+2.74) | 5.28   |
| Data2vec 800 (29h) | 92.25 | 91.06 | 92.5 | 92.69 | 91.94 | 91.56 | 92.88 | 91.5 | 92.25 | 91.31 | 91.99(+0.01) | 8.01   |
| MOCO-v3 500 (48h) | 95.38 | 94.56 | 95.94 | 95.94 | 95.81 | 95.38 | 95.62 | 95.25 | 95.81 | 95.56 | 95.52(+3.54) | 4.48   |
| MAE 800 (34h) | 94.81 | 94.56 | 95.94 | 95.94 | 95.81 | 95.38 | 95.62 | 95.25 | 95.81 | 95.56 | 95.52(+3.54) | 4.48   |
| DAMA 500 (35h) | 94.88 | 94.12 | 95.62 | 95.31 | 95.25 | 95.12 | 94.81 | 94.62 | 95.12 | 94.44 | 94.92(+2.94) | 5.08   |

Table 5. Comparisons of fine-tuning results with state-of-the-art SSL methods pretrained with on Noisy dataset and randomly initialized in accuracy and error rate. Our DAMA reports stable results over dataset settings compared with other state-of-the-arts. We report results of our DAMA and MAE [17] with masking ratios 80% and 60% for Data2Vec [2]. Training epochs and training times are listed along with the methods. Bold and underlined are the highest and second highest scores, respectively.

| Config          | Value            | Config          | Value            |
|-----------------|------------------|-----------------|------------------|
| image size      | 128×128×7        | image size      | 128×128×7        |
| patch size      | 16×16            | patch size      | 16×16            |
| batch size      | 512              | batch size      | 512              |
| epochs          | 500              | epochs          | 150              |
| optimizer       | Adam             | optimizer       | Adam             |
| base learning rate | 1.5e-04       | Base learning rate | 1e-02           |
| min learning rate | 0              | min learning rate | 1e-5            |
| weight decay    | 0.05             | weight decay    | 0.05             |
| learning rate schedule | cosine decay | learning rate schedule | cosine decay |
| warmup epochs   | 40               | warmup epochs   | 5                |
| augmentation    | RandomResizedCrop | augmentation    | RandomResizedCrop |
| K-blocks/β     | 6/2              | K-blocks/β     | 0.1/0.25/0.8/1.0 |

Table 6. Pretraining (left) and finetune (right) setting of our DAMA.
Figure 6. Visualization segmentation validation set of DAMA and other methods at threshold IoU = 0.75. By focusing more on the contextual information, DAMA detects and segments cells better where cells are dense and overlap on each other, e.g., cluster of cell.
Figure 7. Bounding box overall-all-all Precision-Recall curve of DAMA and other SSL methods.

Figure 8. Segmentation mask overall-all-all Precision-Recall curve of DAMA and other SSL methods. Same as Fig. 2.
Figure 9. Bounding box Precision-Recall curve at different IoU threshold of DAMA and other SSL methods. DAMA has the best scores at the IoU from $0.1 : 0.75$ and are competitive numbers at $0.8 : 0.9$.

Figure 10. Segmentation mask Precision-Recall curve at different IoU threshold of DAMA and other SSL methods. DAMA has the best scores at the IoU from $0.1 : 0.75$ and are competitive numbers at $0.8 : 0.9$. 
Figure 11. Example of five cell types: microglia, neurons, oligodendrocytes, endothelial, and astrocytes correspond to five biomarkers: Iba1, NeuN, Olig2, RECA1, S100.