MAESTER: Masked Autoencoder Guided Segmentation at Pixel Resolution for Accurate, Self-Supervised Subcellular Structure Recognition

Ronald Xie[1,2,3,4,*,†] Kuan Pang[1,4,*] Gary D. Bader[1,3,4,‡] Bo Wang[1,2,3,‡]

[1]University of Toronto, [2]Vector Institute, [3]University Health Network, [4]The Donnelly Centre

{ronald.xie, kuan.pang, gary.bader}@mail.utoronto.ca , bowang@vectorinstitute.ai

Abstract

Accurate segmentation of cellular images remains an elusive task due to the intrinsic variability in morphology of biological structures. Complete manual segmentation is unfeasible for large datasets, and while supervised methods have been proposed to automate segmentation, they often rely on manually generated ground truths which are especially challenging and time consuming to generate in biology due to the requirement of domain expertise. Furthermore, these methods have limited generalization capacity, requiring additional manual labels to be generated for each dataset and use case. We introduce MAESTER (Masked AutoEncoder guided SegmenTation at pixEl Resolution), a self-supervised method for accurate, subcellular structure segmentation at pixel resolution. MAESTER treats segmentation as a representation learning and clustering problem. Specifically, MAESTER learns semantically meaningful token representations of multi-pixel image patches while simultaneously maintaining a sufficiently large field of view for contextual learning. We also develop a cover-and-stride inference strategy to achieve pixel-level subcellular structure segmentation. We evaluated MAESTER on a publicly available volumetric electron microscopy (VEM) dataset of primary mouse pancreatic islets β cells and achieved upwards of 29.1% improvement over state-of-the-art under the same evaluation criteria. Furthermore, our results are competitive against supervised methods trained on the same tasks, closing the gap between self-supervised and supervised approaches. MAESTER shows promise for alleviating the critical bottleneck of ground truth generation for imaging related data analysis and thereby greatly increasing the rate of biological discovery.

Code available at https://github.com/bowang-lab/MAESTER

1. Introduction

Imaging is widely used in biology to study the organization, morphology, and function of cells and subcellular structures [13, 26, 28, 31, 32]. Segmentation of structures and objects of interest in the acquired images is often critical for downstream analysis. Recent innovations in high throughput imaging technology enables larger scale datasets to be collected more quickly and cost efficiently [23, 24, 32]. Scalable and accurate segmentation hence becomes a crucial bottleneck to overcome. For example, volumetric electron microscopy (VEM) can generate terabytes of imaging data in a single run, enabling biologists to uncover ultrastructural features of cells at unprecedented resolution and scale in 3D [24]. Manual segmentation of such datasets are unfeasible and especially when substantial domain knowledge is required for annotation of structures captured in the imaging volume.

With recent advancements in the field of machine learning, automatic methods involving convolutional neural networks (CNNs) have been developed to aid the segmentation process to great success [8, 18]. However, these methods often require extensive manual labels to train in the first place. Furthermore, supervised models often exhibit limited generalization capacity, necessitating additional ground truth generation efforts for each new dataset or use case. Presently, there is a dire need for a self-supervised segmentation method to bypass the initial bottleneck of manual label generation, particularly when the cost and time of acquiring training supervision far exceeds the capacity to generate unlabelled data.

In addition to being self-supervised, the method needs to incorporate a few inductive biases to tackle the challenges of biological image segmentation. First, the texture of objects from the same class often remains consistent, despite great variability in shapes and sizes that cellular structures can exhibit. Therefore, the model needs to learn semantically meaningful representation of small image patches belonging to each structure of interest and distinguish between different textures. Second, the model needs to be capable of producing

[*] Equal contribution
[†] Project lead
[‡] Co-senior author

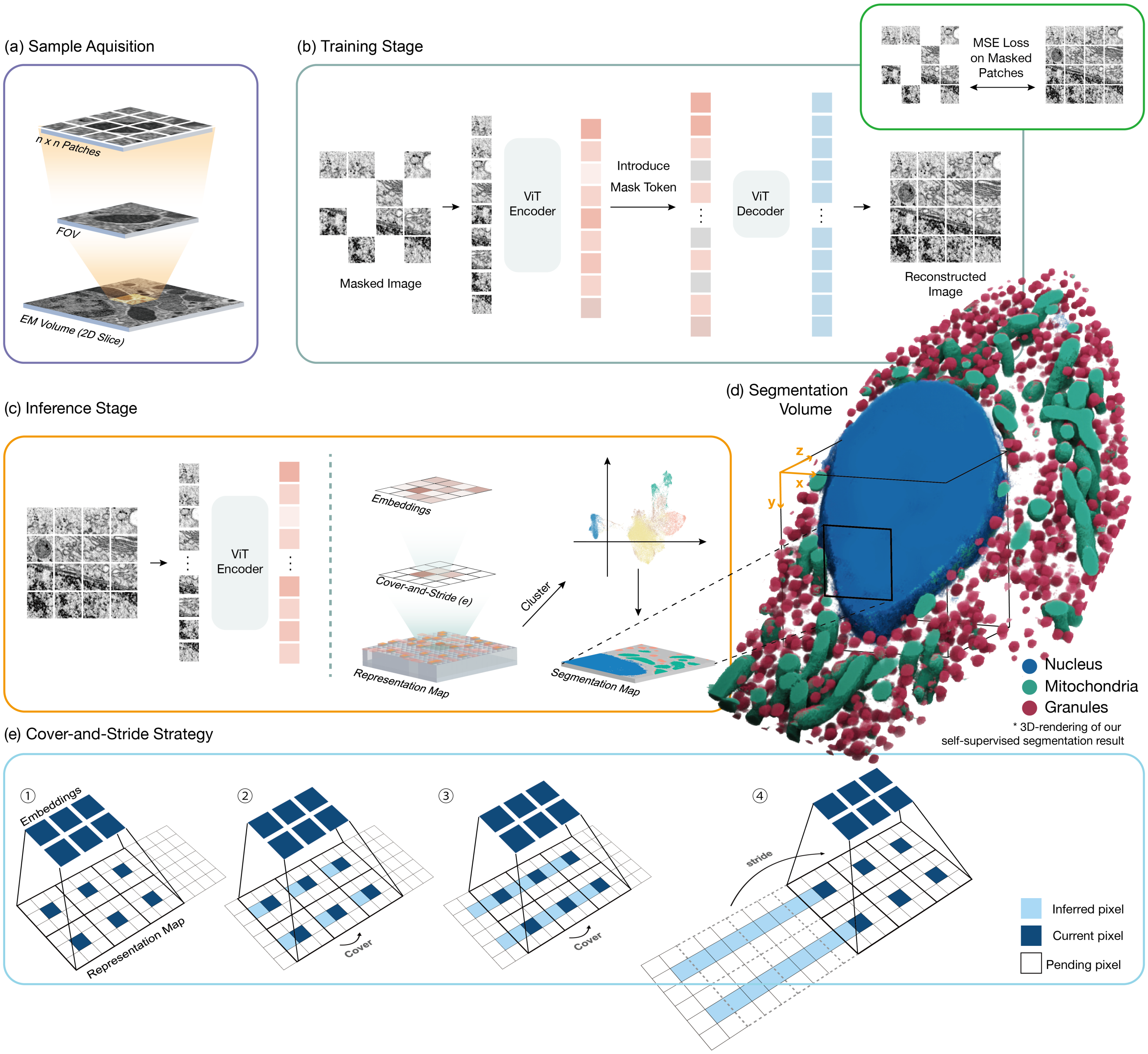

Figure 1. MAESTER achieves self-supervised representation learning and segmentation through: (a) patchifying large sample from EM imaging (b) learning patch-level representation through predicting randomly masked region, (c) inferring the representation for the center voxel of each patch, (d) showing the 3D-rendered volume of our MAESTER generated segmentation, (e) demonstrating cover-and-stride strategy.

features that precisely correspond to small regions in the original image. Not only will this increase the resolution of the resulting segmentation, it will also allow the method to take advantage of the locality assumption, which posits that small groups of adjacent pixels are more likely to belong to the same class. Third, the model needs to be context aware. While distinguishing between multi-pixel patches of images alone can achieve good segmentation results [5], we hypothesize that including a greater field of view (FOV) as context is crucial for better representation learning for the purpose of subcellular structure segmentation.

Transformer based architectures have seen recent successes in computer vision [3, 15, 29]. The token-wise representation of image patches offers a natural way to inject inductive bias into our self-supervised segmentation model. We introduce MAESTER (Masked AutoEncoder guided SegmenTation at pixEl Resolution), a self-supervised method that can achieve accurate, pixel-level segmentation of subcellular structures. MAESTER works in two stages. During training, MAESTER takes as input a large FOV ($F \times F$ pixels) containing ample local context which is further broken down into multi-pixel patches of size $P \times P$ pixels. The choice of P is sufficiently small to allow each patch to be treated as a single class under the locality assumption while achieving higher spatial precision. The attention mechanism of a vision transformer (ViT) [3] encoder then allows information sharing between nearby patches. Furthermore, taking inspiration from the Masked Autoencoder (MAE) [6] learning paradigm, we incorporate the surrogate task of multi-pixel patch masking and reconstruction via a light weight ViT decoder for each sampled FOV of a given image to simultaneously learn semantically meaningful token representations of all patches in the FOV. During inference, we deploy the trained encoder to generate millions of representations of unlabelled image patches via a novel cover-and-stride inference strategy. These representations are then clustered to produce a desired number of classes for self-supervised segmentation, leading to the final segmentation of the given VEM dataset.

To our knowledge, we are the first to use the transformer architecture to incorporate the inductive biases needed for self-supervised subcellular structure segmentation. We also repurposed and optimized the MAE learning paradigm for generating semantically relevant token representations of multi-pixel sized image patches for classification into biologically concordant clusters for segmentation rather than for pretraining or representation learning at the image level. Lastly, we introduce a cover-and-stride inference strategy to achieve pixel level segmentation of the given biological images. We tested MAESTER on the betaSeg dataset [20], consisting of primary mouse pancreatic islets β cells and yielded upwards of 29.1% increase in performance compared to prior state-of-the-art [5]. We also benchmarked against Segmenter [27] and vanilla ViT [3], two supervised segmentation models with access to all ground truth labels in addition to the raw images used to train MAESTER. We find MAESTER achieved competitive results for the predominant classes, closing the gap between supervised and self-supervised segmentation models. We believe MAESTER has the potential to drastically speed up the experimental cycle of biological imaging experiments by alleviating the critical bottleneck of manual label generation and greatly increasing the rate of scientific inquiry in cell biology.

2. Related Work

2.1. Self-supervised representation learning

Self-supervised representation learning in computer vision is a powerful way to learn and extract semantically meaningful features from unlabelled images. This is especially important in the context of biology as manual labels often require extensive domain knowledge and time to generate. Popular approaches for self-supervised learning involves the optimization of one or more surrogate tasks. These could include reconstruction, denoising, color augmentation and rotation prediction [11,22,30,34]. More recently, contrastive learning and diffusion based approaches achieved great results on a number of vision benchmarks [1,4,7,9]. However, many of these approaches were explored for the purpose of self-supervised pretraining, in contrast to our work which directly leverages the semantically meaningful representations that the models generate. Furthermore, our work focuses on pixel level representation learning rather than generating representations of entire images.

2.2. Transformer and masked autoencoder framework in vision

The transformer is a powerful and expressive neural network architecture built via the stacking of attention modules and multilayer perceptrons (MLP) [29]. Transformer in combination with self-supervised learning has been widely used in the field of natural language processing (NLP). For example, BERT [2] was trained via the random masking and prediction of words in sequence as a form of masked language modeling (MLM). More recently, the transformer architecture had also shown promise in the field of computer vision (e.g. Vision transformer(ViT) [3, 15]) as consecutive image patches can be converted into token embeddings synonymous to words in a sentence.

The masked autoencoder (MAE) is a novel self-supervised representation learning paradigm [6]. It has achieved great results in self-supervised pretraining and image level representation learning. For our work, we repurposed and optimized this framework to generate semantically relevant token representations of multi-pixel patches and subsequently grouped them into biologically concordant clusters for segmentation. Importantly, unlike CNNs and many contrastive approaches, the MAE framework is able to support the representation learning of extremely small image patches to achieve precision while simultaneously maintaining a sufficiently large field of view for contextual learning and representation accuracy. This naturally makes MAE an extremely suitable candidate for self-supervised image segmentation in biology.

2.3. Biological image segmentation as a representation learning problem

Segmentation of objects contained in a given image without manual labels remains a challenging task. One approach is to divide the image into small enough patches where one can assume that the entire patch belongs to one of the existing classes. This effectively turns the segmentation problem into a multi-class classification problem given features extracted from individual patches. Several existing methods adopt this mindset to achieve self-supervised segmentation such as JULE (microCT) and MoCo (MRI, CT) [7,14,19,33]. However, these approaches use output prediction as supervisory signal to iteratively refine model prediction, which may lead to the unwanted compounding of model biases and mispredictions. Another study by Han et al. used a variational autoencoder and triplet loss to produce segmentations at the pixel level [5]. However, dividing images into patches as small as 8×8 pixels removes important local context which we demonstrate is necessary for faithful representation learning particularly in the context of biology where concepts such as shape, co-localization and structural heterogeneity etc. span multiple patches but are informative for the representation learning task.

3. Method

3.1. Transformer backbone for precise, context aware representation learning

Convolutional neural networks (CNNs) have long been the predominant backbone of choice for many computer vision tasks [10, 21, 25]. However, for the specific task of representation learning of image patches for segmentation, we argue that CNNs are inferior to transformers because it is difficult for CNNs to simultaneously generate multiple representations for spatially distinct subsections of the original input image. As a consequence, CNNs are limited to either choose a small field of view to generate representations of each individual patch but give up greater local context, or to choose a larger field of view to capture local context, but give up the precision of the learned representation and greatly increase inference time.

Recent works demonstrated the promise of using transformers in vision for many classification tasks [3, 15]. Here, we also demonstrate that the transformer architecture is a natural way to achieve high precision and fidelity of learned patch representations for the purpose of self-supervised segmentation in biology. Specifically, we choose a small patch size ($P = 5$ pixel) to learn spatially precise token representations of image patches while simultaneously maintaining a sufficiently large field of view ($F = 80$ pixel). The attention mechanism and positional embedding intrinsic to the transformer enables context aware learning, producing better token representations of each individual image patch of the given field of view in parallel for a total of 16x16 (n) representations generated per input field of view.

3.2. Learning semantically meaningful token representations of image patches

As shown in Figure 1, we use Vision Transformer(ViT) [3] to implement our strategy of fusing local and large-scale features, and we introduce a surrogate task of patch reconstruction inspired by the Masked Autoencoder(MAE) learning paradigm to extract semantically meaningful representations of each image patch. [6]

During the training stage, FOVs of size $F \times F$ are randomly sampled from the training image volume which then gets divided into non-overlapping $n \times n$ multi-pixel patches. Then we randomly mask out some patches according to the selected masking ratio. The unmasked patches are then injected with positional embedding before being input into the ViT encoder. The resulting embeddings of visible patches are then passed to a separate, light weight ViT decoder where learnable mask tokens are introduced as placeholders for the original masked image patches for patch reconstruction. The ViT decoder reconstructs masked out image patches by predicting the pixel intensities. Following MAE [6], our model is trained via the mean squared error (MSE) loss between the reconstructed image and the original image on masked patches $\frac{1}{D}\sum_{i=1}^{D}(x_i - y_i)^2$ where D is the total number of reconstructed pixels in the FOV.

3.3. Generating pixel level segmentation during inference via cover-and-stride and clustering

During inference, the input FOVs are patchified and fed into the trained ViT encoder without the random masking operation which generates token representations of image patches ($n \times n$) in parallel. To achieve pixel level representation over the entire image, we introduce a cover-and-stride inference strategy as shown in Figure 1(e). Due to the choice of P being sufficiently small relative to the sizes of the biological structures, we were able to treat the resulting token representations as a close proxy to the representation of the center pixel of each multi-pixel patch to effectively increase the resolution of the segmentation. To prevent potential edge bias (i.e. when the patches away from the center of the field of view are not exposed to sufficient local context), we only store the representation generated from a fixed number of patches ($m \times m$) near the FOV center to generate the final pixel level representation map. m is equal to 4 under our default settings, hence 16 token representations are generated and stored simultaneously per input FOV. In order to accommodate this, cover-and-stride first takes single-pixel strides to *cover* a given FOV center, producing a fully inferred area with side length $m \times P$, followed by bigger *strides* of size $m \times P$ to go over the entire image. This process generates and stores pixel level representations in parallel, resulting in over 600 million pixel-wise token representations on the betaSeg testing dataset. Due to practical considerations, we randomly subset $500,000$ patches to compute k-mean [16] centers for unsupervised label assignment, generating a pixel level segmentation of the original image. We visualize the resulting clusters via UMAP [17] using a further subset of $50,000$ image patches to demonstrate semantic separation

config	ViT encoder	ViT decoder
embedding dimension	192	128
transformer layer	14	7
attention head	1	8
MLP ratio	2.0	2.0
positional embedding weight	0.08	1.0

Table 1. MAESTER implementation details under default settings.

of the learnt token representations into putative classes.

3.4. Method details and design decisions

Our choice of the encoder and decoder is loosely based on the ViT-B architecture [3] while making some modifications tailored to our specific purpose of representation learning for segmentation. The default implementation parameters are detailed in Table 1. Our modifications include: 1) limiting the embedding dimensions to compensate for the smaller vocabulary size of multipixel patches compared to larger image patches, while still maintaining sufficient expressivity, 2) further reducing the depth and capacity of the decoder to promotes more information to be encoded in the representation rather than the decoder and 3) introducing a weighted positional embedding in the encoder to dampen its contribution to the representations, preventing biased clustering based on their relative positions in the FOV. More details are discussed in 4.2 and supplementary materials.

4. Experiment

4.1. Data

OpenOrganelle [8] is an public collection of high-resolution cell imaging datasets. Following existing work, we tested our method on a primary mouse pancreatic islet β cell dataset named "BetaSeg" in OpenOrganelle [8, 20]. The dataset was acquired via Focused Ion Beam Scanning Electron Microscopy (FIB-SEM) on two pancreatic tissue samples. Two groups were treated respectively with high-dosage or low-dosage of glucose. We chose the high-dosage group for comparison with existing works. Preprocessing involved cropping cells from the tissue stack into separated volumes and downsampling the resolution from 4 nm to 16 nm. The resulting dataset contains four cell volumes and paired reference segmentations for each cell. The reference segmentations were generated by human annotators or via manual corrections of deep learning models [20].

Each reference segmentation includes binary segmentation masks of 7 subcellular structures, namely centrioles, nucleus, plasma membrane, microtubules, golgi body, granules and mitochondria. We classify the remaining regions of the cell into the unrecognized category. Thus, we have 8 total classes as the reference segmentation. Among these 8 categories, nucleus, granules, mitochondria and unrecognized are the predominant classes. For evaluation, we train our model on cell 1, 2, and 3. We hold out cell 4 for an independent test set.

4.2. Hyperparameters

We chose an FOV size (F) of 80×80 pixels and a patch size (P) of 5×5 pixels, effectively patchifying the input FOV into 16×16 (n) multi-pixel patches. The physical size of our FOV corresponds to a 1.28 micrometers squared area. We chose this setting empirically based on our ablation results, which is in accordance with biological intuition as the subcellular structures of interest are typically 1 micrometer to 5 micrometer in size. At a field view of 1.28 micrometers, our model can successfully capture object level information while our small patch size enables spatially precise token representations to be generated while also staying true to the locality assumption of one class per patch. We learn a 192 dimension representation vector for each multi-pixel patch. We adjust the positional encoding weight to 0.08 in the ViT encoder to avoid positional bias in clustering. We keep the central 4×4 multi-pixel patches in the representation map. We apply masking ratio at 0.5 for MAE to learn biological context, discussed next.

4.3. Reconstruction analysis

Contrary to natural images which typically have multiple channels and a large spatial redundancy, high resolution electron microscopy images have only one channel and seem to have significantly greater information density. We found experimentally that using a high masking ratio such as 0.85 is not adequate for our specific use case. Considering this, we set the masking ratio to 0.5 to retain sufficient information for learning and reconstruction of multi-pixel patches of a given FOV while simultaneously preventing the reconstruction task becoming too trivial. As seen in Figure 2, our method successfully reconstructs various cellular structures, even when substantial information is removed from the original input image. This suggests that the representations learned by the model are semantically meaningful, and likely encode higher level concepts that span multiple patches such as shape and co-localization to inform nearby patch reconstruction.

4.4. Evaluations

Since our model captures more implicit structure that is not listed as an individual category in reference segmentation, we perform a class merge to the prediction for fair comparison. For example, our method detects transparent vesicles as a distinct class, which is classified into the uncategorized category in the reference segmentation. Therefore, we merge the prediction of transparent vesicles into uncategorized category for evaluation.

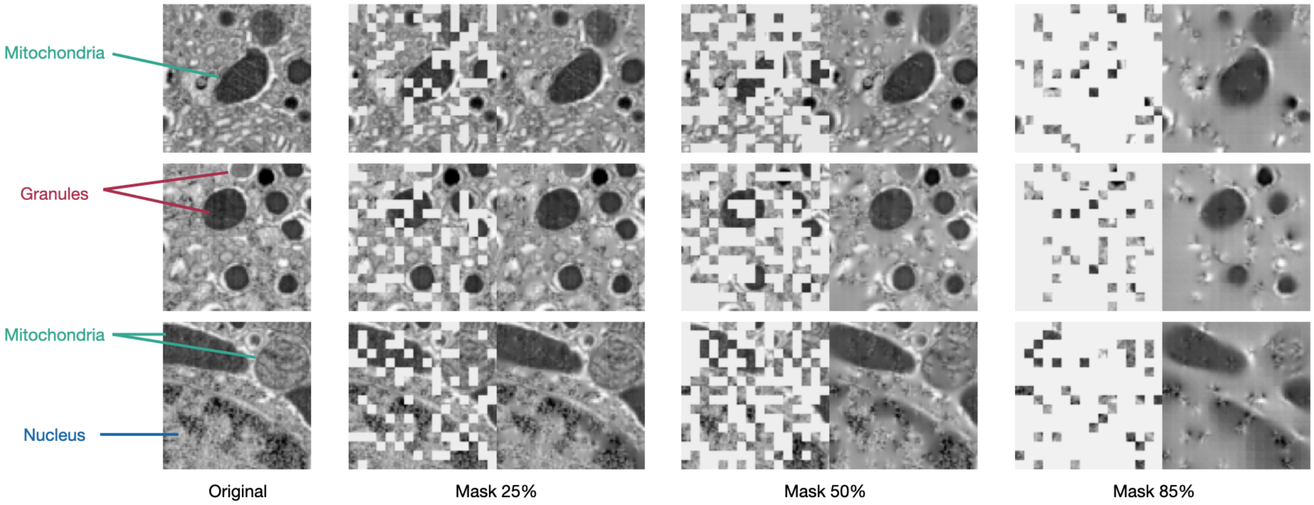

Figure 2. Reconstruction results on the betaSeg testing dataset using a model trained at 50% masking ratio. The reconstructions were performed at different masking ratios over 3 representative FOVs, demonstrating generalizability and encoding quality of the trained model.

We then generate a confusion matrix on the reference segmentation versus our prediction $C_{K_{Pr} \times K_{RS}}$ where K_{Pr} and K_{RS} are the number of classes of the prediction and reference segmentations respectively. Following previous work, we use the most predominant four classes namely the nucleus, mitochondria, granules and unrecognized classes ($K_{RS} = 4$) in the reference segmentation and vary K_{Pr} between 4-10 prior to class merging. We evaluate model performance via the Dice Similarity Coefficient (DSC) based on the resulting confusion matrix after class merging. We use the Hungarian algorithm [12] to match the prediction labels to the lowest cost category in the reference segmentation, also following conventions of prior work [5].

4.5. Self-supervised segmentation

MAESTER is able to generate self-supervised segmentation of subcellular structures without any expert annotation. As shown in Figure 3, the model archives precise, pixel level segmentation on the betaSeg testing dataset. Details regarding architectures and implementation can be found in Supplementary Materials. MAESTER is able to accurately classify the predominant subcellular structures in betaSeg, namely nucleus, granules, mitochondria, and the unrecognized category, which is mostly cytosol. It is worth noting that the prior state-of-the-art method [5] often failed to distinguish between nucleus and cytosol. This is likely because these two classes share textural similarities that could look identical for a given multi-pixel patch. However, our model is able to overcome this by incorporating greater local context to distinguish between differences in arrangement of these textural patterns and utilize object-level landmarks to separate these two classes semantically in representation space. The confusion matrix in Supplementary Figure 2 also reflects this point.

Quantitatively, our method consistently outperforms previous work under the same evaluation setting. We test our methods on k-means centers from 4 to 10. As shown in Table 2, we achieve performance improvements ranging from 11.4% to 29.1% across different K.

4.6. Comparison with supervised baselines

We compare MAESTER ($K = 6$) with Segmenter [27] and Vanilla ViT [3], two supervised baselines with full access to paired ground truth labels in addition to the raw images used to train MAESTER and find that the resulting DSC was somewhat competitive on 3 out of 4 major classes. While there are more mispredictions by MAESTER, the differences in performance could in part be attributed to the inclusion vs. exclusion of boundary pixels in the reference segmentation. Furthermore, due to the lack of domain knowledge, the remaining granules class, which consists of a darker center and a surrounding white membrane bound region, was near impossible to segment correctly as they were regarded as semantically distinct by MAESTER. Representations of the white portion of the granules are closer to background than to its darker granule counterparts. Considering this, MAESTER closes the gap between supervised and self-supervised segmentation models, especially when compared with prior state of the art self-supervised segmentation methods.

4.7. Ablations

From the result of various ablation studies summarized in Table 4, we make a few interesting observations. Results in 4a shows that larger FOV seems to always improve segmentation performance given constant patch size, further

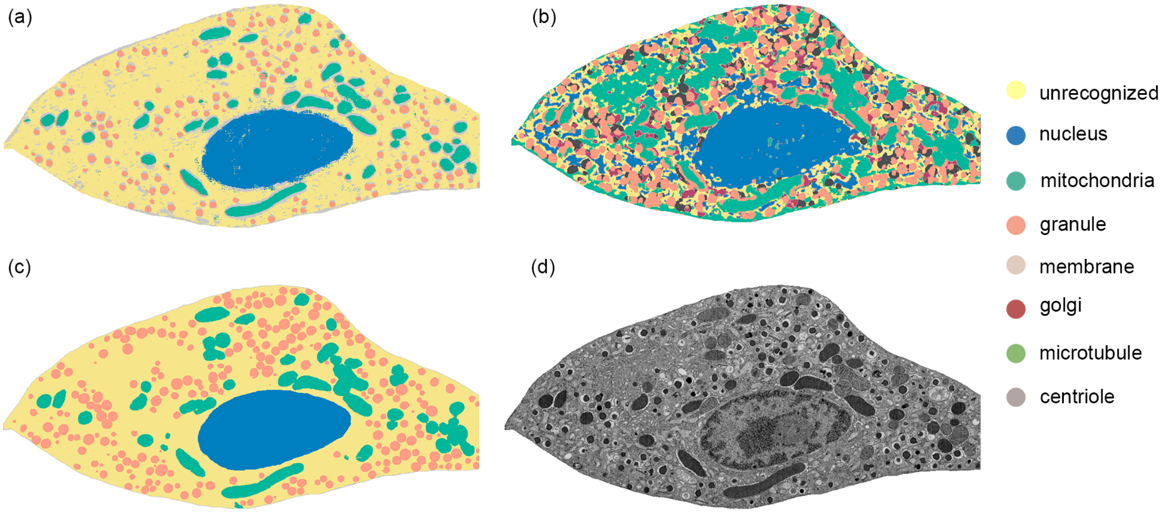

Figure 3. Qualitative results on slice 627 of test cell stack. (a) Segmentation result generated by our method, (b) segmentation result adapted from Han et al. [5] with color alignment, (c) reference segmentation, (d) raw image.

K	Han et al.	MAESTER (Ours)	↑ Improvement
4	0.625	**0.696**	11.4%
5	0.659	**0.773**	17.3%
6	0.647	**0.787**	21.6%
7	0.643	**0.793**	23.3%
8	0.560	**0.723**	29.1%
9	0.578	**0.739**	27.9%
10	0.567	**0.708**	24.9%

Table 2. DSC results comparison on the betaSeg testing dataset across different number of K-means centers. $4 \leq K \leq 10$. Han et al. results are taken from the original manuscript [5].

Class	MAESTER	Segmenter	Vanilla ViT
	(Ours)	(Supervised)	(Supervised)
nucleus	0.950	**0.990**	0.981
granules	0.556	**0.860**	0.774
mitochondria	0.786	**0.896**	0.868
unrecognized	0.844	**0.912**	0.907

Table 3. DSC by class of MAESTER with $K = 6$ compared with Segmenter [27] and Vanilla ViT [3], two supervised baselines with complete access to paired ground truth labels in addition to all the raw images used to train MAESTER.

providing evidence that local context is important for better representation learning. 4b shows that masking ratio is important for defining the balance between usefulness and feasibility of the reconstruction task. A higher masking ratio makes learning difficult when too much non-redundant information is masked out whereas a low masking ratio trivializes learning which results in non-informative representations. 4c demonstrates a speed-accuracy trade-off where keeping representations of patches closer to the center of the FOV generates better segmentation at the cost of inference speed. 4d depicts the trade off between precision and variability of patches of different sizes. Smaller patches contain too little information while larger patches no longer learn accurate representations of the center pixel for inference, losing spatial precision. 4e demonstrates that position conscious contextual learning is important for model performance, supporting our claim that MAESTER is learning and using higher level concepts such as shape and co-localization to achieve better segmentation performance. Lastly, 4f demonstrates the performance boost afforded by our novel cover-and-stride strategy. In particular, it shows that the multi-pixel patch level representation we learn is a good proxy for pixel-level representation of the center pixel in the patch for sufficiently small patches.

5. Discussions and Conclusion

In this work, we present MAESTER, Masked AutoEncoder guided SegmenTation at pixEl Resolution for biological images. MAESTER is capable of generating accurate, pixel-level segmentations of subcellular structures in biological images, demonstrating its potential in alleviating the critical bottleneck of manual ground truth generation in image related biological experiments. MAESTER achieves this by learning semantically meaningful, context aware token representations of multi-pixel patches of a given image and through a new cover-and-stride inference strategy and subsequent clustering, generating the final segmentation result.

Much of the improvement over prior state-of-the-art methods was made possible because we repurposed and optimized the masked autoencoder (MAE) learning paradigm for self-

	8×8	12×12	16×16
K	patches	patches	patches
4	0.601	0.664	**0.696**
5	0.531	0.667	**0.773**
6	0.621	0.666	**0.787**
7	0.635	0.697	**0.793**
8	0.557	0.679	**0.723**
9	0.604	0.695	**0.739**
10	0.578	0.644	**0.708**

(a) **FOV**. Larger FOV improves the segmentation quality when the patch size is kept constant.

K	0.25	0.5	0.85
4	0.671	0.696	**0.710**
5	0.697	0.773	**0.801**
6	0.685	**0.787**	0.721
7	0.667	**0.793**	0.722
8	0.687	**0.723**	0.694
9	0.673	**0.739**	0.691
10	0.676	**0.708**	0.658

(b) **Masking Ratio**. Masking ratio is important for learning biologically relevant token representations.

K	4×4	6×6	8×8
4	**0.696**	0.696	0.696
5	**0.773**	0.738	0.734
6	**0.787**	0.753	0.750
7	**0.793**	0.682	0.624
8	**0.723**	0.666	0.682
9	**0.739**	0.690	0.669
10	**0.708**	0.698	0.707

(c) **Number of kept tokens ($m \times m$) per FOV during inference**. Accuracy of segmentation improves further when m decreases. However, this is a trade off between efficiency and accuracy.

K	3×3 px	5×5 px	7×7 px	9×9 px
4	0.439	0.696	**0.800**	0.745
5	0.405	**0.773**	0.740	0.705
6	0.371	**0.787**	0.731	0.688
7	0.341	**0.793**	0.708	0.672
8	0.320	**0.723**	0.706	0.664
9	0.315	**0.739**	0.700	0.643
10	0.307	**0.708**	0.696	0.636

(d) **Patch Size**. Smaller patch sizes contain too little information while larger patch sizes violate the locality assumption.

	w/	w/o
K	PosEmbed	PosEmbed
4	**0.696**	0.614
5	**0.773**	0.633
6	**0.787**	0.647
7	**0.793**	0.625
8	**0.723**	0.586
9	**0.739**	0.601
10	**0.708**	0.608

(e) **Positional Embedding**. Position-conscious contextual learning is important for model performance.

K	Stride Only	C & S
4	0.608	**0.696**
5	0.715	**0.773**
6	0.720	**0.787**
7	0.665	**0.793**
8	0.713	**0.723**
9	0.717	**0.739**
10	0.700	**0.708**

(f) **Cover-and-Stride**. Cover-and-Stride (C&S) strategy improves segmentation precision.

Table 4. MAESTER ablation experiments on test cell stack. We report DSC for $4 \leq K \leq 10$. Default settings are marked in gray.

supervised segmentation rather than its original purpose of pretraining or image-level representation learning. By drastically reducing the patch size to only a few pixels, we learn spatially precise token representations of multi-pixel image patches while staying true to the locality assumption which allows us to more appropriately frame unsupervised segmentation as a classification task compared to other methods following a similar mindset. It is worth noting that without spatial context and information sharing between patches, selecting an extremely small patch size is not possible in the first place due to the resulting lack of information contained in individual multi-pixel patches.

It is also worth highlighting that although MAESTER achieves high quality segmentation compared to the reference segmentation, we make the observation that the reference segmentation and current evaluation metrics are not adequate for assessing MAESTER's true capabilities in subcellular structure recognition. For example, MAESTER separately classifies the outer membranes of organelles in contrast to the reference segmentation, which groups boundary pixels with the unrecognized class, causing a drop in our model performance. As inter-organelle membranes are naturally delineated, MAESTER has the potential to achieve unsupervised instance segmentation of organelle classes.

Nonetheless, due to the nature of self-supervised approaches, MAESTER does suffer from the lack of domain knowledge. For example, granules of β cells are a class of organelle consisting of two high contrast components - a darker center and a surrounding white membrane bound region. Without any biological background, MAESTER distinguishes these two semantically different areas as two separate classes, which contributes to the gap in performance compared to supervised methods. Therefore, in the future, we are interested in representing and incorporating domain knowledge into our self-supervised segmentation algorithm to further increase performance.

On the other hand, as the self-supervised segmentation results of MAESTER are competitive against supervised baselines for many predominant classes, we are also interested to use the resulting segmentation as a weakly supervised signal to train or finetune other models to further speed up the experimental cycles of imaging experiments in biology.

6. Acknowledgements

All authors thank the Vector Institute, Calcul Québec, and the Digital Research Alliance of Canada for their support. This research has been made possible in part by a grant to G.D.B. from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation. B.W. is supported by the NSERC discovery grant and CIFAR chair programs [RGPIN-2020-06189, DGECR-2020-00294]. R.X. is supported by the Ontario Graduate Scholarship.

References

[1] 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. 3

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, June 2019. 3

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2, 3, 4, 5, 6, 7

[4] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 3

[5] Hongqing Han, Mariia Dmitrieva, Alexander Sauer, Ka Ho Tam, and Jens Rittscher. Self-supervised voxel-level representation rediscovers subcellular structures in volume electron microscopy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1874–1883, 2022. 2, 3, 4, 6, 7

[6] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 3, 4

[7] 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. 3, 4

[8] Larissa Heinrich, Davis Bennett, David Ackerman, Woohyun Park, John Bogovic, Nils Eckstein, Alyson Petruncio, Jody Clements, Song Pang, C Shan Xu, et al. Whole-cell organelle segmentation in volume electron microscopy. *Nature*, 599(7883):141–146, 2021. 1, 5

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 3

[10] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021. 4

[11] Nikos Komodakis and Spyros Gidaris. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*, 2018. 3

[12] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 6

[13] Jeff W Lichtman and José-Angel Conchello. Fluorescence microscopy. *Nature methods*, 2(12):910–919, 2005. 1

[14] Lihao Liu, Angelica I Avilés-Rivero, and Carola-Bibiane Schönlieb. Contrastive registration for unsupervised medical image segmentation. *arXiv preprint arXiv:2011.08894*, 2020. 4

[15] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2, 3, 4

[16] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982. 4

[17] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 4

[18] Manca Žerovnik Mekuč, Ciril Bohak, Samo Hudoklin, Byeong Hak Kim, Min Young Kim, Matija Marolt, et al. Automatic segmentation of mitochondria and endolysosomes in volumetric electron microscopy data. *Computers in biology and medicine*, 119:103693, 2020. 1

[19] Takayasu Moriya, Holger R Roth, Shota Nakamura, Hirohisa Oda, Kai Nagara, Masahiro Oda, and Kensaku Mori. Unsupervised segmentation of 3d medical images based on clustering and deep representation learning. In *Medical Imaging 2018: Biomedical Applications in Molecular, Structural, and Functional Imaging*, volume 10578, pages 483–489. SPIE, 2018. 4

[20] Andreas Müller, Deborah Schmidt, C Shan Xu, Song Pang, Joyson Verner D'Costa, Susanne Kretschmar, Carla Münster, Thomas Kurth, Florian Jug, Martin Weigert, et al. 3d fib-sem reconstruction of microtubule–organelle interaction in whole primary mouse β cells. *Journal of Cell Biology*, 220(2), 2021. 3, 5

[21] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015. 4

[22] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 3

[23] Christopher J Peddie and Lucy M Collinson. Exploring the third dimension: volume electron microscopy comes of age. *Micron*, 61:9–19, 2014. 1

[24] Christopher J Peddie, Christel Genoud, Anna Kreshuk, Kimberly Meechan, Kristina D Micheva, Kedar Narayan, Constantin Pape, Robert G Parton, Nicole L Schieber, Yannick Schwab, et al. Volume electron microscopy. *Nature Reviews Methods Primers*, 2(1):1–23, 2022. 1

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing*

and computer-assisted intervention, pages 234–241. Springer, 2015. 4

[26] David J Stephens and Victoria J Allan. Light microscopy techniques for live cell imaging. *science*, 300(5616):82–86, 2003. 1

[27] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021. 3, 6, 7

[28] M Titford. The long history of hematoxylin. *Biotechnic & histochemistry*, 80(2):73–78, 2005. 1

[29] 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. 2, 3

[30] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008. 3

[31] Daniel Witvliet, Ben Mulcahy, James K Mitchell, Yaron Meirovitch, Daniel R Berger, Yuelong Wu, Yufang Liu, Wan Xian Koh, Rajeev Parvathala, Douglas Holmyard, et al. Connectomes across development reveal principles of brain maturation. *Nature*, 596(7871):257–261, 2021. 1

[32] C Shan Xu, Kenneth J Hayworth, Zhiyuan Lu, Patricia Grob, Ahmed M Hassan, José G García-Cerdán, Krishna K Niyogi, Eva Nogales, Richard J Weinberg, and Harald F Hess. Enhanced fib-sem systems for large-volume 3d imaging. *elife*, 6:e25916, 2017. 1

[33] Jianwei Yang, Devi Parikh, and Dhruv Batra. Joint unsupervised learning of deep representations and image clusters. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5147–5156, 2016. 4

[34] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 3