Cross-patch Dense Contrastive Learning for Semi-supervised Segmentation of Cellular Nuclei in Histopathologic Images

Huisi Wu[1], Zhaoze Wang[1], Youyi Song[2], Lin Yang[2], Jing Qin[2]
[1]Shenzhen University, [2]The Hong Kong Polytechnic University

Abstract

We study the semi-supervised learning problem, using a few labeled data and a large amount of unlabeled data to train the network, by developing a cross-patch dense contrastive learning framework, to segment cellular nuclei in histopathologic images. This task is motivated by the expensive burden on collecting labeled data for histopathologic image segmentation tasks. The key idea of our method is to align features of teacher and student networks, sampled from cross-image in both patch- and pixel-levels, for enforcing the intra-class compactness and inter-class separability of features that as we shown is helpful for extracting valuable knowledge from unlabeled data. We also design a novel optimization framework that combines consistency regularization and entropy minimization techniques, showing good property in eviction of gradient vanishing. We assess the proposed method on two publicly available datasets, and obtain positive results on extensive experiments, outperforming the state-of-the-art methods. Codes are available at https://github.com/zzw-szu/CDCL.

1. Introduction

Deep learning models have achieved remarkable success in cellular nuclei segmentation from histopathologic images [9, 42, 43]. However, for a good learning performance, we often have to collect a large amount of annotated data that tell how the deep model should output. The crux is such annotated data are rather time-consuming to collect or can be even prohibitively expensive, because they need tedious efforts from domain experts and the annotating process have to be conducted multiple rounds for reaching a consensus among experts. Therefore, it has been receiving an increasing interest for researchers to study on how to train deep models given only a few annotated data.

Seminal works in this direction include mainly semi-supervised learning [17, 21, 36] and weakly-supervised learning [1, 18, 37]. In this work, we study semi-supervised learning problem that just needs a few pixel-wise annotated data while being able to learn from quite massive unlabeled data. This setting is more suitable for segmenting cellular nuclei in histopathologic images where a lot of objects presented in one image, thus weak supervision signals, say object's center, often being far more expensive to acquire.

Advanced semi-supervised learning techniques in medical image segmentation often are based on adversarial training, pseudo-labeling, and consistency regularization [10, 23, 26]. These existing methods, though have shown being able to leverage knowledge from unlabeled data for learning, suffer from lack of exploiting feature structures across the whole dataset, such as the similarity or disparity exists between different features. Our idea for solving this problem is to tie semi-supervised learning and contrastive learning together. Contrastive learning selects positive and negative pairs of features from unlabeled data, and then exploit knowledge from them by contrasting the similar features against dissimilar features, being able to learn high-level semantic structures across different images.

The key to implement our idea is the sampling quality of positive and negative pairs. Existing methods are based on pixel-wise sampling, positive pairs consisting of multiple views of perturbations from the same pixel-wise feature while negative pairs are randomly sampled by features with different pixel-wise predictions, under the guidance of pseudo labels [20, 44]. Since nuclei have blur boundaries while obvious distributions, pseudo labels are not accurate as expected in pixel-wise, though they still reflect the class distribution in a fixed region (e.g., patch), suggesting that it is easier to correctly judge inter-patch feature disparity rather than inter-pixel. Therefore, by leveraging the inter-patch feature disparity, deep models are more likely to learn better representations of target distributions.

With the above insight, we develop a cross-patch dense contrastive learning framework to extract structural information from unlabeled data. Specifically, we sample patch-wise negative pairs between patches with large disparity and densely sample pixel-wise negative pairs between them. Following with the standard positive sampling strategy, our contrastive learning module enforces the intra-class compactness and inter-class separability [38] in both patch-level and pixel-level. With the contrastive learning mod-

ule as the core, we further take the advantage of the mean teacher architecture, consistency regularization, and entropy minimization to acquire predictions and pseudo labels with higher quality, for the semi-supervised segmentation of cellular nuclei in histopathologic images. We conduct extensive experiments on two publicly available datasets, with the positive results showing the effectiveness of our method, consistently outperforming the state-of-the-art methods. We summarize the contributions as follows.

- We propose an effective and generic cross-patch dense contrastive learning framework to extract valuable knowledge from unlabeled data, by enforcing the intra-class compactness and inter-class separability in both patch-level and pixel-level.
- We take the advantage of consistency regularization and entropy minimization, develop an efficient semi-supervised nuclei image segmentation algorithm that outperforms the state-of-art methods in two publicly available datasets.

2. Related Work

2.1. Cellular Nuclei Segmentation in Histopathologic Images

Cellular nuclei segmentation is a preliminary but complicated task in computer-assisted diagnosis and tumor microenvironment analysis [42]. Traditional techniques often use background subtraction and color thresholding [27, 29] that need complex post-processing to provide segmentation results, and so they are unable to handle challenging cases such as overlap and occlusion in nuclei images. With the advance of CNN, deep learning models have been extensively applied to the task of nuclei segmentation, while most of them have achieved high accuracy only in the fully supervised settings [24, 32, 45]. However, limited annotation data hinders the generalizability of the existing nuclei segmentation approaches. Therefore, it is urgent to develop methods that can be trained with limited supervision and extract information from unlabeled data.

2.2. Semi-supervised Semantic Segmentation

Segmentation methods based on semi-supervised learning have been shown to be able to address the aforementioned problem by exploiting information from unlabeled data. For example, adversarial training methods [26, 35] utilize generative adversarial network [12] to extract useful structural information from unlabeled data. Pseudo-labeling methods [8, 10] create artificial labels for unlabeled data by retaining model predictions with high confidence. Other mainstream works take advantage of unlabeled data by enforcing a consistency over different perturbations. TCSMv2 [23] adopts the self-ensembling architecture and enforces a transformation-consistency to improve the performance of the output-level regularization. CutMix [11] encourages a mixture-consistency between the mixed predictions and predictions generated by mixed inputs. GCT [16] proposes a detector to approximate pixel-wise prediction confidence with a dynamic consistency constraint. CCT [30] takes the outputs of the encoder as the object of perturbation, which can enhance the network's ability of representation learning by preserving the invariance of the predictions over different perturbations. Different from them, we propose a novel semi-supervised segmentation method which demonstrates the superiority of integrating self-supervised contrastive learning.

2.3. Contrastive Learning

Contrastive learning is a highly regarded technique for learning representations from unlabeled features these days [6, 7, 14]. It aims to obtain better representation learning by contrasting similar features (positive pairs) against dissimilar features (negative pairs). An important innovation direction for contrastive learning is how to select positive/negative pairs. Besides, memory bank is adopted to store more negative samples since they can lead to better performance [6]. In the semantic segmentation field, there are lots of works that leverage contrastive learning for the pre-training of models [4, 39, 41]. But recently, Wang *et al.* [38] has shown the advantages of applying contrastive learning in a cross-image pixel-wise manner for supervised segmentation. CAC [20] demonstrates its improvement in semi-supervised segmentation by performing directional contrastive learning pixel-to-pixel to align lower quality feature towards its counterpart. Following these works, we propose a cross-patch dense contrastive learning module as the core of our semi-supervised segmentation method.

3. Method

The proposed semi-supervised segmentation method, as illustrated in Figure 1, is based on the mean teacher framework [36]. The student and teacher models share the same architecture, consisting of an extractor, a classifier, and a projector. The supervised branch (black arrows in Figure 1) exploits labeled data by calculating L_{sup}, the standard cross-entropy loss, between predictions and ground truths. In the unsupervised branch, contrastive learning on projector output features with a contrastive loss L_{contr}, as well as consistency regularization and entropy minimization on classifier output predictions with losses L_{cons} and L_{ent}, drive the network extract information from unlabeled data.

The student model is optimized by a weighted summation of the above losses, formulated as:

$$L = w_{sup}L_{sup} + w_{contr}L_{contr} + w_{cons}L_{cons} + w_{ent}L_{ent}, \quad (1)$$

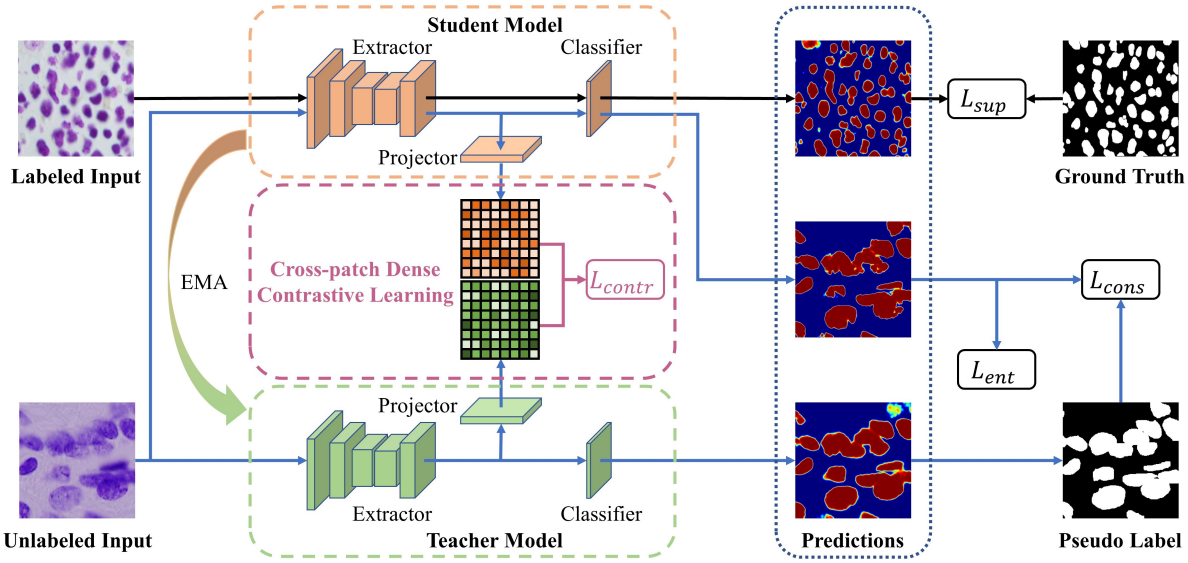

Figure 1. The overall architecture. The student and teacher models share the same architecture. We update student model by reducing a weighted summation of L_{sup}, L_{contr}, L_{cons} and L_{ent}. Teacher model is updated by setting as an EMA of weights of the student model. The black arrows represent the supervised branch, while the others represent the unsupervised branch.

where w are weighted factors used to balance the impact of individual loss terms. We update teacher model by setting as an exponential moving average (EMA) of weights of the student model, rather than the commonly used gradient descent technique. The procedure can be formulated as:

$$\theta_i^t = \alpha\theta_{i-1}^t + (1-\alpha)\theta_i^s, \tag{2}$$

where θ_i^t stands for teacher model's weights at i-th iteration, while θ_i^s for those of the student model. The $\alpha \in [0, 1]$ is a balance weight for the updating. Note that above temporal ensembling of weights can help the teacher yield more accurate predictions [36], which facilitate the unsupervised training of the student model and eventually optimize the segmentation results.

3.1. Cross-patch Dense Contrastive Learning

The distribution of cellular nuclei in histopathologic images is generally scattered. Hence, we can divide the image into multiple fixed-size patches, each of which contains different proportions of foreground and background pixels. Considering it is relatively easy to judge the inter-patch feature disparity with the assistance of pseudo labels, our cross-patch dense contrastive learning module, as shown in Figure 2, is developed based on the idea of original pixel-wise contrastive learning and patch-wise contrastive learning. Following with previous works [20, 38], this module consists of two stages: positive/negative pairs sampling and contrastive loss calculation.

Cross-patch Dense Sampling. An overview of the proposed sampling strategy is shown in Figure 2. Strongly and weakly augmented inputs of the same image are passed into the student and teacher model respectively. The extractor outputs are projected into low dimension feature maps, where we sample positive/negative pairs in both patch-level and pixel-level. The projector can preserve the crucial contextual information in the extracted features, which has been proved to be beneficial for contrastive learning [6].

Following with the standard positive sampling strategy, we select a patch-wise feature from student model and its positive counterpart is sampled from the corresponding place in the teacher model. Between these two patches, pixel-wise features with the same position form pixel-wise positive pairs. Our strategy differs in negative sampling. We consider two patch-wise features with large disparity as a negative pair and then sample pixel-wise pairs in a cross-patch dense manner. To measure inter-patch feature disparity, we introduce a patch-wise metric, which is calculated based on pseudo labels.

Specifically, student and teacher models' pseudo labels $\widetilde{y}_u^s$, $\widetilde{y}_u^t$ are obtained as follows:

$$\widetilde{y}_u^s = argmax(P_u^s), \tag{3}$$

$$\widetilde{y}_u^t = argmax(P_u^t), \tag{4}$$

where P_u^s and P_u^t represent student and teacher models' predictions of unlabeled data, and the metric, call foreground score FS, is calculated as follows:

$$FS = \frac{N_f}{N}, \tag{5}$$

$$N_f = \sum_{h,w} \mathbb{1}\{\widetilde{y}_u^{h,w} = 1\}, \tag{6}$$

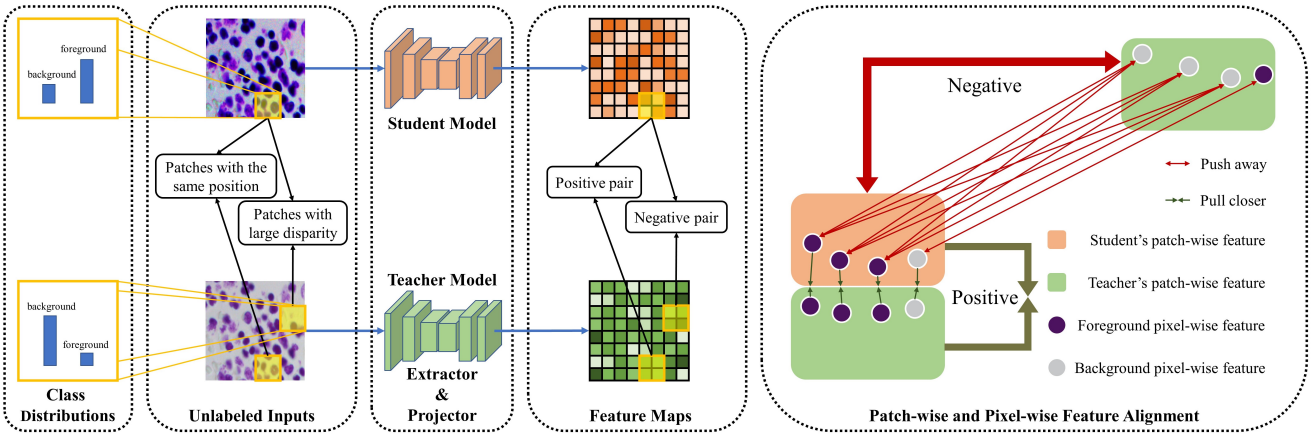

Figure 2. An illustrative pipeline of the proposed cross-patch dense contrastive learning for semi-supervised nuclei segmentation. Patch-wise features with large disparity are sampled as patch-wise negative pairs, while pixel-wise negative pairs are densely sampled between them. Positive pairs are sampled between patches and pixels with the same position, and the module pulls positive pairs closer and pushes negative pairs away in both patch-level and pixel-level.

where N is the total number of pixels and N_f is the number of foreground pixels in a selected patch. The $\hat{y}_u$ is equivalent to $\hat{y}_u^s$ or $\hat{y}_u^t$, depending on the patch acquired from the student model or the teacher model. FS indicates the proportion of pixels belong to the target class. According to FS, we divide these patches into three categories: FDPs (Foreground-dominate Patches, $FS \geq 0.7$), BDPs (Background-dominate Patches, $FS \leq 0.3$) and MPs (Mixed Patches, $0.3 < FS < 0.7$). Patches structurally dissimilar in label space should also have dissimilar distributions in feature space, and hence two patch-wise features corresponding to FDP and BDP respectively have large disparity and can be sampled as a patch-wise negative pair.

Since FS only focus on the class proportion while ignoring the spatial distribution, one-to-one pixel-wise sampling between two patches can only obtain limited effective pixel-wise negative pairs. Therefore, we adopt a many-to-many approach to densely take the pixels between two patches with large feature disparity as pixel-wise negative pairs, which can provide stronger constraint to increase inter-class separability. Note that we leverage pseudo labels again to filter false negative pairs. Furthermore, since increasing negative pairs can enhance contrastive learning [6], we maintain two feature banks (BDB and FDB, which represent Background-dominate Bank and Foreground-dominate Bank) to store patch-wise features processed in the last few iterations to guarantee adequate negative counterparts.

Pixel-wise and Implicit Patch-wise Contrastive Loss. After sampling positive/negative pairs in both patch-level and pixel-level, we design the contrastive loss to pull positive pairs closer and push negative pairs away. We first formulate the contrastive loss function for a certain query pixel-wise feature q, modified on the basis of InfoNCE [28]:

$$l_{contr}(q) = -log \frac{sim(q, k_+)}{sim(q, k_+) + \sum\limits_{k_- \in FB} \mathcal{J}_{q,k_-} sim(q, k_-)} \tag{7}$$

$$sim(q, k) = exp\left(\frac{q^T k}{\|q\| \, \|k\| \, \tau}\right) \tag{8}$$

$$\mathcal{J}_{q,k_-} = \mathbb{1}\{\widetilde{y}_q \neq \widetilde{y}_{k_-}\} \tag{9}$$

where sim represents the exponential equation of cosine similarity and τ the temperature. The k_+ and k_- denote the positive and negative counterparts for q, respectively. FB represents the feature bank and its type is determined by the source of q. If q is from a FDP, FB is the BDB, and vice versa. The $\mathcal{J}_{q,k_-}$ is a binary mask defined for judging whether the pseudo labels for the two features in a negative pair are different, and based on it, the false negative pairs are discarded.

Given the student and teacher models' feature maps, $\mathcal{F}_s$ and $\mathcal{F}_t$, we define Φ_s as a patch-wise feature in $\mathcal{F}_s$ and Φ_t as the location-corresponding one in $\mathcal{F}_t$. Based on l_{contr}, the pixel-wise contrastive loss for Φ_s is defined as:

$$l_{contr}^{\Phi}(\Phi_s) = \frac{1}{N} \sum_{h,w} l_{contr}(\phi_s^{h,w}), \tag{10}$$

where $\phi^{h,w}$ denotes the pixel-wise feature with spatial locations h and w in Φ, N denotes the number of pixels in the patch, and $\phi_t^{h,w}$ is the only positive counterpart for $\phi_s^{h,w}$.

By minimizing l_{contr}^{Φ}, the network learns to contrast pixel-wise features from the same class against those from different classes, so we can obtain better predictions in detail. We then implicitly formulate our patch-wise contrastive learning in the unified pixel-wise total loss for $\mathcal{F}_s$, which can be formulated as follows:

$$L_{contr} = l^{\mathcal{F}}_{contr}(\mathcal{F}_s) = \frac{1}{N^{B,F}_s} \sum_{i=1}^{N^{B,F}_s} l^{\Phi}_{contr}(\Phi^i_s), \quad (11)$$

$$N^{B,F}_s = \sum_{i=1}^{N^{\Phi}_s} \mathbb{1}\{\Phi^i_s \in BDP \; or \; \Phi^i_s \in FDP\}, \quad (12)$$

where N^{Φ}_s denotes the number of patches in $\mathcal{F}_s$ and $N^{B,F}_s$ is the total number of BDPs and FDPs, whose internal class distribution is quite imbalance. In this case, we calculate $l^{\mathcal{F}}_{contr}$ as the average of losses on every BDP and FDP. By minimizing $l^{\mathcal{F}}_{contr}$, the network learns to contrast structurally similar patch-wise features against those with large disparity, driving the model towards better predictions in the target distribution. In addition, $l^{\mathcal{F}}_{contr}$ is the final contrastive loss for current training image (or batch, just let $\mathcal{F}$ represent a group of feature maps). In this regard, we extend the level of loss calculation from former "pixel-image/batch" to "pixel-patch-image/batch", aiming at better exploiting feature structures across these histopathologic nuclei images.

3.2. Consistency Regularization

Though contrastive learning on the intermediate feature maps effectively learns strong feature representations from unlabeled data, it often fails to directly optimize the parameters of the classifier. Inspired by FixMatch [34] and its applications in segmentation [2, 40], we introduce a simplified consistency regularization on the segmentation predictions to overcome this shortcoming. Considering the teacher model provides more accurate and robust predictions [36], we set teacher's prediction as target and let the student model converge to it. For every teacher's prediction P^t_u, we compute the pseudo label $\hat{y}^t_u$ as follows:

$$\hat{y}^t_u = argmax(P^t_u). \quad (13)$$

The consistency loss L_{cons} for unlabeled data is calculated by the cross-entropy:

$$L_{cons} = \mathcal{H}(P^s_u, \hat{y}^t_u), \quad (14)$$

where P^s_u is the student's prediction and $\mathcal{H}$ is the cross-entropy loss function.

3.3. Entropy Minimization

The predictions are confidence maps indicating the probability of a pixel belong to each class. They can be exploited to produce pseudo labels, which are harnessed to guide the positive/negative pairs sampling and play an important role in our contrastive learning module.

Following with entropy minimization [13] and its applications in segmentation [5, 25], we introduce a regularization loss calculated on student's prediction P^s_u, which is formulated as:

$$L_{ent} = -\frac{1}{N} \sum_{n=1}^{N} \sum_{c=1}^{C} P^{s,n,c}_u log P^{s,n,c}_u, \quad (15)$$

where N and C represent the numbers of pixels and classes. While consistency regularization and entropy minimization improve the correctness of predictions, contrastive learning obtains more reliable pseudo-labeling guidance and better performance, which conversely optimizes the confidence of predictions. The combination of L_{contr}, L_{cons} and L_{ent} aims at promoting this virtuous circle.

4. Experiments

4.1. Datasets

We assess our method on two publicly available datasets, denoted by DSB and MoNuSeg, obtained from the 2018 Data Science Bowl challenge [3] and multi-organ nuclei segmentation challenge [19], respectively.

DSB Dataset. This dataset includes 670 nuclei images from different modalities of brightfield and fluorescence, where the target boundary is also difficult to identify.

MoNuSeg Dataset. This dataset consists of a training set with 30 histopathologic images and a test set with 14 images, all of which are H&E stained tissue images from multi organs, where the low contrast exists between targets and background tissues.

4.2. Implementation Details

Network Architecture. We use DenseUNet [22] as the base segmentation network and DenseNet-161 [15] as the backbone, pretrained on ImageNet [33]. The extractor in Figure 1 refers to all other components except the final classifier in DenseUNet, with 256 output channels. The projector is implemented by a $FC \rightarrow ReLU \rightarrow FC$ architecture and reduces the number of channels to 128.

Hyperparameter Settings. For an input image with size $h \times w$, the patch size for contrastive learning is set to $\frac{h}{8} \times \frac{w}{8}$ pixels in image space and its corresponding patch-wise feature is $\frac{h}{64} \times \frac{w}{64}$ pixels. To obtain a better trade-off between memory usage and performance of contrastive learning, we adjust the two feature banks to store patch-wise features from the current and previous batches. Specifically, we use the gradient checkpoint function imported from PyTorch [31] to prevent the oversized feature bank from significantly increasing the training burden. The different loss weights for L_{sup}, L_{contr} and L_{ent} are set as fixed values, which is as follows: $w_{sup} = 1$, $w_{contr} = 0.1$, $w_{ent} = 0.01$. w_{cons} grows from 0 to 1 along a Gaussian curve with the

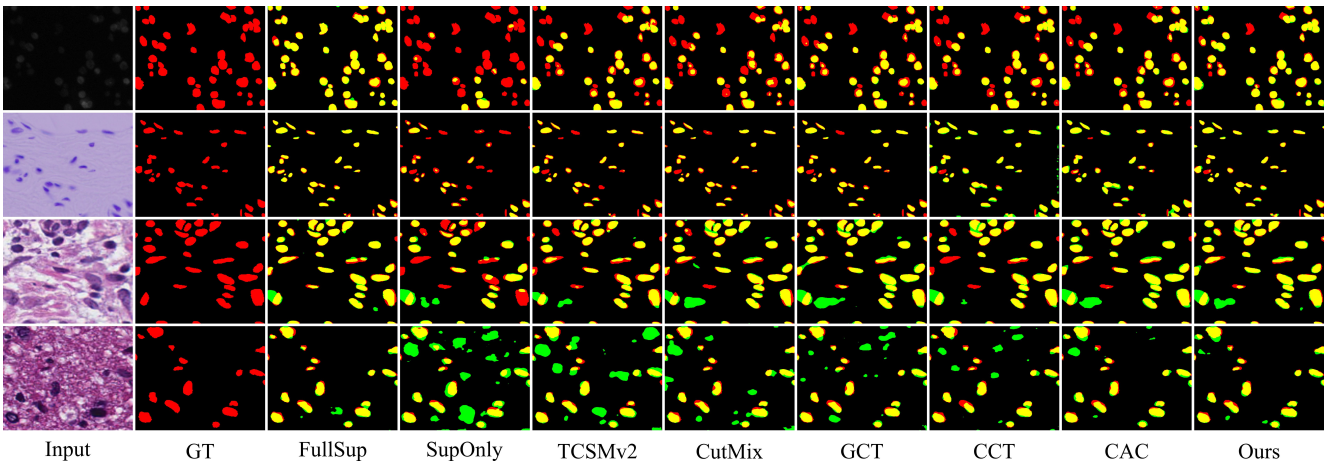

Figure 3. Visual comparison with different state-of-the-art methods in nuclei image segmentation. FullSup is trained with 100% labeled data while SupOnly with only 1/32. Their results represent upper and lower bounds of other methods. Other methods are trained in a semi-supervised manner with 1/32 labeled data and 31/32 unlabeled data. Green and red pixels indicate the predictions and ground truth respectively. Yellow pixels represent the overlap regions between the prediction and ground truth.

increase of training epoch. The unsupervised branch does not start training until the 6th epoch, to obtain a better initialization of the network and reduce the impact of deviating pseudo labels. Following the most effective configuration, we set α, the weight for EMA, to 0.999. Based on the poly learning rate scheduling strategy, our initial learning rate is set to 0.0001. For a fast convergence, we also employ an Adam strategy as our optimizer in the training process. By setting batch size of both labeled and unlabeled images as 8, our network usually can be converged within 80 epochs.

Data Augmentation. Due to the limited image data, we conduct 6 kinds of data augmentation techniques to alleviate overfitting; they are (1) random flipping, (2) random cropping, (3) random rotation with a degree in [-15, 15], (4) random Gaussian blur, (5) color jitters, and (6) gray scaling.

4.3. Comparison with State-of-the-art Methods

We compare our method against several state-of-the-art methods, including TCSMv2 [23], CutMix [11], GCT [16], CCT [30] and CAC [20]. We implement all competitors with the same base segmentation network, as well as the same experimental environments and data augmentations, to ensure the fairness of comparison.

To sufficiently demonstrate the effectiveness of our method, we randomly divide the DSB dataset into 3 parts with ratio 7:1:2, which are used for training, validation, and test respectively. For MoNuSeg dataset, we also randomly select 20% of its public training set for validation. Since the size of MoNuSeg's images is 1000×1000 pixels, we uniformly crop sub-images of size 250×250 with no overlap from these images. All images from DSB and sub-images from MoNuSeg are augmented and resized to a uniform resolution of 320×320 pixels as the inputs of model training.

We perform a statistical comparison with state-of-the-art methods by collecting the Dice coefficient (DC), Jaccard coefficient (JC), accuracy (ACC), specificity (SP) and sensitivity (SE) over the DSB and MoNuSeg datasets, in which DC and JC are the main indicators to measure precision of biomedical segmentation. From the results shown in Table 1, we can clearly see that our method generally outperforms other competitors in all settings with different amounts of labeled images, including 1/32, 1/16 and 1/8 of the total training images. Especially with 1/32 labeled training data, our method surpasses CAC by 1.09% on DSB, 1.18% on MoNuSeg, in DC metric.

Figure 3 presents the visual comparison with the FullSup, SupOnly and state-of-the-art methods. We observe that, against the competitors, our method obtains better predictions of both cellular nucleus' distributions (number, location) and details (shape, size), while also comparable to the FullSup method, reflected by less over-predicting (green pixels) and under-predicting (red pixels) in our results.

4.4. Ablation Studies

To assess each component of the proposed method, we perform the following ablation studies on DSB and MoNuSeg dataset, with only 1/32 of the training data being labeled.

Without any semi-supervised technique, DenseUNet [22], the base segmentation network, can only leverage labeled images for training (SupOnly). By sampling pixel-wise positive/negative pairs and calculating L_{contr}, pixel-wise contrastive learning is accomplished, for helping the network learn in a semi-supervised manner (Scheme.1, Scheme.2). Our cross-patch dense contrastive learning module takes advantage of inter-patch feature disparity by sampling both pixel- and patch-wise pairs

DSB							MoNuSeg					
Label	Method	DC(%)	JC(%)	ACC(%)	SP(%)	SE(%)	Label	Method	DC(%)	JC(%)	ACC(%)	SP(%)	SE(%)
1/32	SupOnly	83.93	73.96	96.12	97.52	85.31	1/32	SupOnly	71.83	56.49	87.47	88.76	82.44
	TCSMv2	85.04	75.40	96.07	96.61	**90.97**		TCSMv2	73.08	58.42	88.18	88.97	85.78
	CutMix	86.03	77.10	96.97	97.82	89.39		CutMix	73.37	58.69	88.16	88.62	86.86
	GCT	85.50	77.68	**97.00**	**98.55**	83.39		GCT	73.80	59.21	88.39	88.82	**86.91**
	CCT	86.13	76.91	96.75	97.38	88.51		CCT	74.29	59.59	**90.18**	**94.17**	74.83
	CAC	86.40	77.58	96.57	98.27	86.27		CAC	74.79	60.28	88.96	89.93	84.75
	Ours	**87.49**	**79.35**	96.99	98.38	88.43		**Ours**	**75.97**	**61.77**	89.83	91.03	85.36
1/16	SupOnly	86.10	77.17	96.81	98.36	85.03	1/16	SupOnly	74.72	60.29	90.23	**93.82**	76.40
	TCSMv2	87.48	78.48	97.03	**99.29**	83.37		TCSMv2	75.78	61.55	90.49	93.22	79.66
	CutMix	87.70	79.92	97.44	98.54	86.27		CutMix	76.20	62.12	90.01	91.52	84.49
	GCT	88.13	80.60	97.51	98.50	88.75		GCT	76.63	62.66	90.45	92.37	83.19
	CCT	88.15	79.69	97.32	98.25	88.89		CCT	76.59	62.45	90.10	91.71	83.41
	CAC	88.49	80.19	97.28	98.44	88.36		CAC	77.12	63.16	90.24	91.50	**85.17**
	Ours	**89.88**	**82.34**	**97.57**	98.64	**90.11**		**Ours**	**77.77**	**64.07**	**91.01**	93.12	83.04
1/8	SupOnly	87.38	79.24	97.09	98.71	86.05	1/8	SupOnly	75.81	61.50	90.26	92.72	80.65
	TCSMv2	88.42	80.40	97.48	98.37	88.48		TCSMv2	77.44	63.52	90.50	92.32	82.60
	CutMix	88.32	80.92	97.69	**99.03**	86.41		CutMix	77.29	63.70	90.64	92.36	84.52
	GCT	88.80	81.10	97.49	98.53	88.31		GCT	77.69	64.08	90.74	92.20	**85.42**
	CCT	89.08	81.25	97.52	98.73	88.61		CCT	77.21	63.25	90.18	91.32	85.40
	CAC	89.37	81.94	97.63	98.34	**90.79**		CAC	78.24	64.54	91.05	93.04	82.82
	Ours	**90.09**	**82.68**	**97.69**	98.68	90.57		**Ours**	**78.93**	**65.56**	**91.44**	**93.34**	84.28
100%	FullSup	90.46	83.27	97.86	98.77	90.71	100%	FullSup	79.97	66.92	92.37	95.69	79.13

Table 1. Statistical comparison with state-of-the-art methods on the test set.

(Scheme.3). Finally, we employ L_{cons} and L_{ent} as auxiliary losses to perform consistency regularization and entropy minimization on the predictions (Scheme.4, Scheme.5, Scheme.6, Ours). Table 2 shows the DC comparison of the above schemes, while Figure 5 presents partial visual results.

Ablation Studies for S_{samp} and L_{contr}. In Scheme.1, we adopt the mean teacher framework and perform simple pixel-wise contrastive learning with a random sampling strategy. Specifically, pixel-wise features from student and teacher model with the same position form positive pairs, while others form negative pairs. The result does not grow as we expect but drop from 83.93% to 81.68%, 71.83% to 69.08%. In Scheme.2, pseudo labels are utilized to guide negative sampling, by selecting features with different pixel-wise predictions as negative pairs. Scheme.2 outperforms Scheme.1, as well as SupOnly by a large margin, proves that (1) contrastive learning effectively extract knowledge from unlabeled data, (2) the sampling quality of positive/negative pairs heavily affects the performance of contrastive learning. Scheme.3 corresponds to our cross-patch dense contrastive learning module, where we improve the sampling strategy and accomplish feature alignment both in pixel- and patch-levels. Since the purpose of the contrastive learning module is to obtain an extractor that can better distinguish classes of features, we also visualize the feature maps extracted with different schemes, as shown in Figure 4. The sharper contrast between target and non-target features, along with the larger increase in DC metric, reflects that such pixel- and patch-wise alignment is feasible and beneficial, also demonstrates the effectiveness of our module. Note that in Scheme.1, 2, anchor pixel-wise features are selected from BDPs and FDPs we obtained, to ensure a fair ablation. FDB and BDB are merged into one feature bank, which means pixel-wise features from the same patch could be sampled as negative pairs, hindering patch-wise feature alignment.

Ablation Studies for L_{cons} and L_{ent}. Consistency regularization is performed on predictions, by calculating L_{cons}, to provide direct parameter optimization for classifier and ensure a better convergence of both student and teacher models. Entropy minimization is applied to the predictions of student model, by calculating L_{ent}, to optimize classification confidence of each pixel, accelerating the virtuous circle of "better pseudo labels$\rightarrow$ better contrastive and consistency learning performance $\rightarrow$ better pseudo labels". After

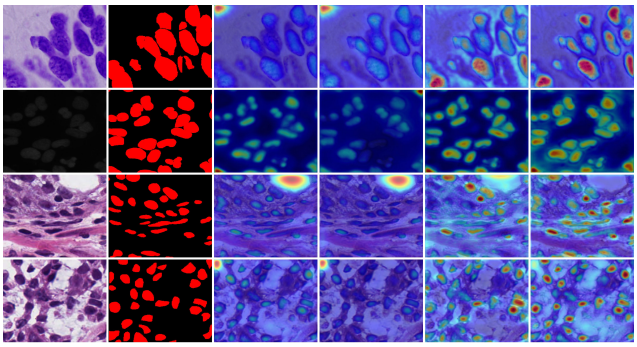

Figure 4. Visual comparison of feature maps extracted with different schemes in ablation studies: (a) Input image, (b) Ground truth, (c) SupOnly, (d) Scheme.1, (e) Scheme.2, (f) Scheme.3.

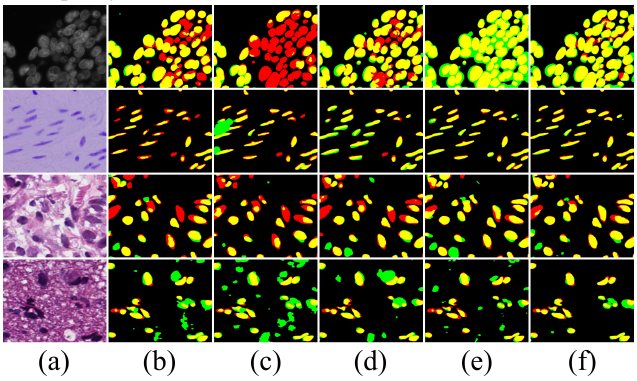

Figure 5. Partial visual comparison of our ablation studies: (a) Input image, (b) SupOnly, (c) Scheme.1, (d) Scheme.2, (e) Scheme.3, (f) Ours. Green and red pixels indicate the predictions and ground truth respectively, while yellow pixels represent their overlap regions.

adding L_{cons} and L_{ent}, the DC metric rises from 86.26% to 87.49%, 74.49% to 75.97%, compared to Scheme.3, which only applies contrastive learning on intermediate features. This also points out it is efficient to take the advantage of consistency regularization and entropy minimization, for complementing contrastive learning and achieving better performance in semi-supervised segmentation. Besides, the combination of only L_{cons} and L_{ent} can also obtain a considerable improvement in the semi-supervised settings, as shown in the result of Scheme.6.

5. Conclusion

We present a novel semi-supervised method for cellular nuclei segmentation in histopathologic images, aiming at efficiently and comprehensively addressing the inherent shortcomings of limited annotated training data. The proposed cross-patch dense contrastive learning module is to accomplish cross-image feature alignment in both patch-level and pixel-level, which enforces the intra-class compactness and inter-class separability over the whole dataset,

Method	S_{samp}	L_{contr}	L_{cons}	L_{ent}	DSB	MoNuSeg
SupOnly					83.93	71.83
Scheme.1	random	✓			81.68	69.08
Scheme.2	pixel	✓			85.54	73.69
Scheme.3	pixel-patch	✓			86.26	74.49
Scheme.4	pixel-patch	✓	✓		86.83	75.19
Scheme.5	pixel-patch	✓		✓	86.65	75.00
Scheme.6			✓	✓	86.40	75.16
Ours	pixel-patch	✓	✓	✓	**87.49**	**75.97**

Table 2. Statistical comparison of our ablation studies in DC metric, with 1/32 labeled training data.

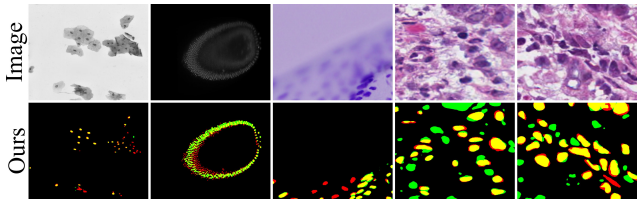

Figure 6. Failure cases for our method. Green and red pixels indicate the predictions and ground truth respectively, while yellow pixels represent their overlap regions.

enabling networks to effectively extract knowledge from unlabeled data. Consistency regularization and entropy minimization are further performed on the network outputs to obtain predictions and pseudo labels with higher quality, which provide guidance for contrastive learning and lead to better segmentation performance.

The above comparative experiments and ablation studies demonstrate the effectiveness of our proposed semi-supervised segmentation method. While our method still fails to segment the cases with extremely small scales, as well as extremely low contrast between the targets and background tissues, as shown in Figure 6, overall, with a very small amount of labeled data, our method handles well a majority of the challenging examples, consistently outperforms the competitors. Future investigations include testing our method on more histopathologic datasets and integrating it in tumor microenvironment analysis systems.

Acknowledgments

This work was supported partly by National Natural Science Foundation of China (No. 61973221), Natural Science Foundation of Guangdong Province, China (Nos. 2018A030313381 and 2019A1515011165), the COVID-19 Prevention Project of Guangdong Province, China (No. 2020KZDZX1174), the Major Project of the New Generation of Artificial Intelligence (No. 2018AAA0102900), and the Hong Kong Innovation and Technology Fund (Project no. ITS/180/20FP).

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2209–2218, 2019. 1

[2] Inigo Alonso, Alberto Sabater, David Ferstl, Luis Montesano, and Ana C Murillo. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8219–8228, 2021. 5

[3] Juan C Caicedo, Allen Goodman, Kyle W Karhohs, Beth A Cimini, Jeanelle Ackerman, Marzieh Haghighi, CherKeng Heng, Tim Becker, Minh Doan, Claire McQuin, et al. Nucleus segmentation across imaging experiments: the 2018 data science bowl. *Nature methods*, 16(12):1247–1253, 2019. 5

[4] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. *arXiv preprint arXiv:2006.10511*, 2020. 2

[5] Minghao Chen, Hongyang Xue, and Deng Cai. Domain adaptation for semantic segmentation with maximum squares loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2090–2099, 2019. 5

[6] 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. 2, 3, 4

[7] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2

[8] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2613–2622, 2021. 2

[9] Yiqi Chen, Xuanya Li, Kai Hu, Zhineng Chen, and Xieping Gao. Nuclei segmentation in histopathology images using rotation equivariant and multi-level feature aggregation neural network. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 549–554. IEEE, 2020. 1

[10] Zhengyang Feng, Qianyu Zhou, Guangliang Cheng, Xin Tan, Jianping Shi, and Lizhuang Ma. Semi-supervised semantic segmentation via dynamic self-training and classbalanced curriculum. *arXiv preprint arXiv:2004.08514*, 1(2):5, 2020. 1, 2

[11] Geoff French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. *arXiv preprint arXiv:1906.01916*, 2019. 2, 6

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[13] Yves Grandvalet, Yoshua Bengio, et al. Semi-supervised learning by entropy minimization. *CAP*, 367:281–296, 2005. 5

[14] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020. 2

[15] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 5

[16] Zhanghan Ke, Di Qiu, Kaican Li, Qiong Yan, and Rynson WH Lau. Guided collaborative training for pixel-wise semi-supervised learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 429–445. Springer, 2020. 2, 6

[17] Zhanghan Ke, Daoye Wang, Qiong Yan, Jimmy Ren, and Rynson WH Lau. Dual student: Breaking the limits of the teacher in semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6728–6736, 2019. 1

[18] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 876–885, 2017. 1

[19] Neeraj Kumar, Ruchika Verma, Deepak Anand, Yanning Zhou, Omer Fahri Onder, Efstratios Tsougenis, Hao Chen, Pheng-Ann Heng, Jiahui Li, Zhiqiang Hu, et al. A multi-organ nucleus segmentation challenge. *IEEE transactions on medical imaging*, 39(5):1380–1391, 2019. 5

[20] Xin Lai, Zhuotao Tian, Li Jiang, Shu Liu, Hengshuang Zhao, Liwei Wang, and Jiaya Jia. Semi-supervised semantic segmentation with directional context-aware consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1205–1214, 2021. 1, 2, 3, 6

[21] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 1

[22] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE transactions on medical imaging*, 37(12):2663–2674, 2018. 5, 6

[23] Xiaomeng Li, Lequan Yu, Hao Chen, Chi-Wing Fu, Lei Xing, and Pheng-Ann Heng. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):523–534, 2020. 1, 2, 6

[24] Dongnan Liu, Donghao Zhang, Yang Song, Chaoyi Zhang, Fan Zhang, Lauren O'Donnell, and Weidong Cai. Nuclei segmentation via a deep panoptic model with semantic feature fusion. In *IJCAI*, pages 861–868, 2019. 2

[25] Weizhe Liu, David Ferstl, Samuel Schulter, Lukas Zebedin, Pascal Fua, and Christian Leistner. Domain adaptation for semantic segmentation via patch-wise contrastive learning. *arXiv preprint arXiv:2104.11056*, 2021. 5

[26] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018. 1, 2

[27] Shivang Naik, Scott Doyle, Shannon Agner, Anant Madabhushi, Michael Feldman, and John Tomaszewski. Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology. In *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 284–287. IEEE, 2008. 2

[28] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

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

[30] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020. 2, 6

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019. 5

[32] Shan E Ahmed Raza, Linda Cheung, Muhammad Shaban, Simon Graham, David Epstein, Stella Pelengaris, Michael Khan, and Nasir M Rajpoot. Micro-net: A unified model for segmentation of various objects in microscopy images. *Medical image analysis*, 52:160–173, 2019. 2

[33] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 5

[34] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020. 5

[35] Jost Tobias Springenberg. Unsupervised and semi-supervised learning with categorical generative adversarial networks. *arXiv preprint arXiv:1511.06390*, 2015. 2

[36] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*, 2017. 1, 2, 3, 5

[37] Paul Vernaza and Manmohan Chandraker. Learning random-walk label propagation for weakly-supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7158–7166, 2017. 1

[38] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. *arXiv preprint arXiv:2101.11939*, 2021. 1, 2, 3

[39] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021. 2

[40] Yicheng Wu, Minfeng Xu, Zongyuan Ge, Jianfei Cai, and Lei Zhang. Semi-supervised left atrium segmentation with mutual consistency training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 297–306. Springer, 2021. 5

[41] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16684–16693, 2021. 2

[42] Bingchao Zhao, Xin Chen, Zhi Li, Zhiwen Yu, Su Yao, Lixu Yan, Yuqian Wang, Zaiyi Liu, Changhong Liang, and Chu Han. Triple u-net: Hematoxylin-aware nuclei segmentation with progressive dense feature aggregation. *Medical Image Analysis*, 65:101786, 2020. 1, 2

[43] Meng Zhao, Hao Wang, Ying Han, Xiaokang Wang, Hong-Ning Dai, Xuguo Sun, Jin Zhang, and Marius Pedersen. Seens: Nuclei segmentation in pap smear images with selective edge enhancement. *Future Generation Computer Systems*, 114:185–194, 2021. 1

[44] Yuanyi Zhong, Bodi Yuan, Hong Wu, Zhiqiang Yuan, Jian Peng, and Yu-Xiong Wang. Pixel contrastive-consistent semi-supervised semantic segmentation. *arXiv preprint arXiv:2108.09025*, 2021. 1

[45] Yanning Zhou, Omer Fahri Onder, Qi Dou, Efstratios Tsougenis, Hao Chen, and Pheng-Ann Heng. Cia-net: Robust nuclei instance segmentation with contour-aware information aggregation. In *International Conference on Information Processing in Medical Imaging*, pages 682–693. Springer, 2019. 2