Efficient Self-Supervision using Patch-based Contrastive Learning for Histopathology Image Segmentation

Boserup, Nicklas; Selvan, Raghavendra

DOI:
10.48550/arXiv.2208.10779

Publication date:
2022

Document license:
CC BY

Citation for published version (APA):
Efficient Self-Supervision using Patch-based Contrastive Learning for Histopathology Image Segmentation

Nicklas Boserup\textsuperscript{1} and Raghavendra Selvan\textsuperscript{1,2,3}

\textsuperscript{1}Dept. of Computer Science, University of Copenhagen
\textsuperscript{2}Dept. of Neuroscience, University of Copenhagen
\textsuperscript{3}Pioneer Centre for AI, Denmark

Abstract

Learning discriminative representations of unlabelled data is a challenging task. Contrastive self-supervised learning provides a framework to learn meaningful representations using learned notions of similarity measures from simple pretext tasks. In this work, we propose a simple and efficient framework for self-supervised image segmentation using contrastive learning on image patches, without using explicit pretext tasks or any further labeled fine-tuning. A fully convolutional neural network (FCNN) is trained in a self-supervised manner to discern features in the input images and obtain confidence maps which capture the network’s belief about the objects belonging to the same class. Positive- and negative- patches are sampled based on the average entropy in the confidence maps for contrastive learning. Convergence is assumed when the information separation between the positive patches is small, and the positive-negative pairs is large. We evaluate this method for the task of segmenting nuclei from multiple histopathology datasets, and show comparable performance with relevant self-supervised and supervised methods. The proposed model only consists of a simple FCNN with 10.8k parameters and requires about 5 minutes to converge on the high resolution microscopy datasets, which is orders of magnitude smaller than the relevant self-supervised methods to attain similar performance.

1 Introduction

Learning task specific representations without any or with limited labelled data continues to be an elusive goal of machine learning. Recent advancements in contrastive learning and self-supervised learning have shown promising results in obtaining discriminative representations of the data which can be useful for downstream applications such as image classification \cite{Chen2020}, object detection \cite{Xie2021}, and speech recognition \cite{Baevski2020}. Contrastive self-supervised learning (CSL) has been successfully used as a form of pre-training to reduce the dependence on labeled data for more complex tasks such as image segmentation \cite{Xie2020}. Most self-supervised methods rely on pretext tasks for training them \cite{Noroozi2016, Sahasrabudhe2020}. Designing relevant pretext tasks can be challenging and even if a useful pretext task is obtained they may not easily generalise across datasets \cite{Misra2020}.

In this work, we present a self-supervised learning framework that contrastively learns an object detection...
model for segmenting nuclei in histopathology images. The model comprises a fully convolutional neural network (FCNN) that predicts one confidence map per output channel, which captures the confidence of each pixel belonging to a particular object class. The FCNN is contrastively trained using smaller positive- and negative-patches stochastically sampled from the images. Patches within a training batch are sampled from an entropy based distribution, where the entropy is based on the patch-level confidence scores. The intuition behind the entropy-based sampling is to obtain positive patches that contain similar information and negative patches that contain contrasting information, with respect to features that can discriminate between objects of different classes. Through iterative training we are able to improve the information separation between the positive- and negative-patches, resulting in an object detection model which can be used for segmentation.

We use a simple FCNN with 10.8k tunable parameters which converges in about 5 minutes on a stand-alone GPU workstation. Experimental evaluation on two histopathology datasets [Kumar et al. 2017, Graham et al. 2019] show that our efficient, contrastive self-supervised learning method obtains performance comparable to relevant supervised and self-supervised baseline methods, using only a fraction of the compute time and resources. An illustration of how the confidence map evolves during the training process is illustrated in Figure 1.

2 Method

Segmentation is fundamentally the task of partitioning an image into areas of interest and a background class. Assuming that the areas of interest within the same class are similar in some feature space according to a similarity measure, we present our contrastive learning framework that results in an unsupervised segmentation model. A high level overview of the proposed framework is illustrated in Figure 2.

2.1 Notation

Consider a batch of \( M \) images, with the \( i \)‘th image represented as the pair \((U_{i}, I_{i})\), where \( U_{i} \subset \mathbb{N}^d \) is a finite set representing the \( d \)-dimensional pixel coordinates and \( I_{i} \) is a function \( I_{i} : U_{i} \rightarrow [0, 1]^\lambda \), with \( \lambda \in \mathbb{N} \) denoting the number of colours/channels, mapping such pixel coordinates into their respective values.

Next, we introduce the notation for denoting the patch locations sampled from the image \( i \) as the tuple \((R, i)\), where \( R \subset U_{i} \) is the set of pixel coordinates of the image patch. Note that the power set \( \mathcal{P}(U_{i}) \) denotes the set of all possible smaller patches that can be sampled from image \( i \). Denoting the set of all such patch tuples across images \( \mathcal{X} \), we consider the subset \( S \subset \mathcal{X} \) of regular (square) and spatially connected patches in this work.

The proposed framework uses a confidence network (Section 2.2), \( f_{\theta} \), parameterised by \( \theta \) which for a given input image \((U_{i}, I_{i})\) computes the confidence map, \( C_{i} \), i.e., \( f_{\theta} : (U_{i}, I_{i}) \rightarrow C_{i} \subset [0,1]^{K} \). The \( k \)‘th coordinate in \( C_{i}^{k}(u) \) indicates the belief, or confidence, that the pixel \( u \in U_{i} \) in image \( i \) belongs to class \( k \), for \( k=1,...,K \), with \( K \geq 1 \).

The contrasting of patches depends on a similarity measure (Section 2.4), \( s : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R} \), which compares the similarity of two patches. This similarity computation relies on \((I_{i}, I_{j})\) and \((C_{i}, C_{j})\), where \( i \) and \( j \) are the images from which the two compared patches originate from.

2.2 Confidence Network

The aim of the confidence network, \( f_{\theta} \), is to learn to detect objects in images without any pixel-level supervision, such that areas with high confidence of the \( k \)‘th class actually belong to a specific type of object in the images. Conversely, objects detected with high confidence of the \( j \)‘th class, when \( k \neq j \), should belong to a different class. The confidence network, \( f_{\theta} \), can be any trainable model which given an input image produces a set of confidence maps, describing the confidence that each pixel of the image belongs to a corresponding class. This, for instance, can be implemented as a fully convolutional neural network (FCNN), with multi-class support i.e., with multiple output channels such that each channel would correspond to a different output class.

The confidence network is trained to discriminate between objects that could belong to different classes by contrasting patches of images against each other. Obtaining meaningful positive- and negative-patches for the contrastive learning is the foundational problem in this framework. We next describe a novel approach to mine for such contrastive patches for self-supervision of the confidence network.
2.3 Entropy-based Patch Sampler

For each output class \(k = 1, \ldots, K\), patches believed to contain (a part of) an object belonging to a specific class are treated as positive patches, whereas patches believed to not contain (a part of) an object of that class are treated as negative patches. Appropriate sampling of such positive- and negative- patches is of vital importance, as the optimisation of the confidence network, \(f_{\theta}\), is performed according to the contrastive loss computed based on these patches. Improved object detection shall ideally correlate with positive and negative sampled patches being more distinct. Effectively, the appropriate sampling of patches becomes a form of pretext/auxiliary task for training the confidence network, \(f_{\theta}\).

Assume a set of candidate patches, \(S \subseteq X\). For each patch \((R, i) \in S\) with \(|R|\) pixels and for each class \(k\), the average patch confidence is computed as:

\[
A_k(R, i) := \frac{1}{|R|} \sum_{u \in R} C^k_i(u) \in [0,1]. \tag{1}
\]

Notice that the higher the average patch confidence, the stronger the belief that the patch contains the type of object belonging to class \(k\).

Recall that \(C_i \rightarrow [0,1]^K\), that is, for each class, \(k = 1, \ldots, K\), a \(C^k_i(u)\) value closer to 1 indicates a higher probability of \(u\) belonging to the \(k\)th class. Likewise, a value closer to 0 indicates that \(u\) likely does not belong to the \(k\)th class, whereas values in between indicate varying degrees of (un)certainty in either direction. Using this intuition about the confidence maps, we model a Bernoulli distributed random variable, \(X(u)\), with confidence value as its parameter, denoted as \(X(u) \sim \text{Bern}(C^k_i(u))\). Using this, we can now define the average patch entropy as

\[
B_k(R, i) := \frac{1}{|R|} \sum_{u \in R} H(X(u)), \text{ where } \tag{2}
\]

\[
H(X) = - \sum_{x \in \{0,1\}} P(X(u) = x) \log_2 P(X(u) = x) \tag{3}
\]

\[
= - C^k_i(u) \log_2 C^k_i(u) - (1 - C^k_i(u)) \log_2 (1 - C^k_i(u)), \tag{4}
\]

which is the entropy of the Bernoulli distributed random variable \(X(u)\). The motivation for this choice of Bernoulli random variable is so that good choices of positive- and negative- patches, according to the contrastive loss, would correlate with higher certainty from the confidence network. Therefore sampling a set \(W_k \subseteq S\) of \(n\) patches according to the unnormalised distribution

\[
1 - B_k(R, i)
\]

for each class \(k\) ensures a stochastic correlation between the confidence map and the appropriateness of patch sampling. Finally, partitioning of these \(W_k\) into sets of positive samples, \(P_k\) and negative samples \(N_k\) is performed according to

\[
[(R, i) \in P_k] \sim \text{Bern}(A_k(R, i))
\]

for each patch \((R, i) \in W_k\), such that patches with high confidence of belonging to class \(k\) are likely to be selected as positive for class \(k\). Additional sampling strategies are described in Appendix B.

2.4 Similarity Measures

Ideally, the similarity measure, \(s:\mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}\), should yield higher values when two input patches are
more similar. In this work, for any two patches \((R_i, v), (R_j, v') \in S\), the similarity measure performs the comparison – not on the input image but – on the image scaled by the corresponding confidence map, given by the following pixel-wise product: 
\[ F_i(u) := I_i(u)C_i(u), \quad u \in R_1, \quad \text{and} \quad F_j(v) := I_j(v)C_j(v), \quad v \in R_2. \]

We employ two pixel-level similarity measures: mean squared error (MSE) and cross-entropy (CE), and treat them as a model hyperparameter and report the best performing measures for each of the datasets.

**Backpropagation:** The scaling of the input image with the confidence map also serves the purpose of connecting the gradients between the sampling step and the confidence network for backpropagating the contrastive loss to optimise the confidence network.

### 2.5 Contrastive Loss

Within a channel \(k\) of the confidence map, we seek to maximise similarity between positive patches while minimising the similarity between positive- and negative- patches. This is facilitated using the intra-channel contrastive loss

\[
L_{\text{intra}} = -\frac{1}{|P_k|^2} \sum_{r,p \in P_k} \log \frac{\exp(s(r,p))}{\sum_{n \in N_k} \exp(s(r,n))}. 
\]  

This contrastive loss, inspired by SimCLR [Chen et al. 2020], rewards \(f_b\) for learning to detect a single feature in the images for each class in the confidence map, when patches are sampled appropriately. However, this lacks any mechanism for enforcing the individual classes to learn distinct features. In a multi-class scenario, each of the classes shall ideally detect different features. This is achieved by introducing an inter-channel contrastive loss

\[
L_{\text{inter}} = -\frac{1}{|P_k|^2} \sum_{r,p \in P_k} \sum_{k=1}^{K} \frac{\exp(s(r,p_k))}{\sum_{k=1}^{K} \sum_{p \in P_k} \exp(s(r,p))}. 
\]  

Finally, the combined loss is defined as

\[
L = L_{\text{intra}} + L_{\text{inter}}. 
\]

### 3 Data & Experiments

#### 3.1 Data

As the proposed CSL framework is trained exclusively using unlabelled images, without exposure to the labels during training, we can also use the labels from the training set for validation of the segmentation performance. That is, we train the weights of the FCNN without using any labels but can employ the labels for model selection.

**MoNuSeg:** This dataset Kumar et al. [2017] from MICCAI 2018 contains 37 training images at 1000 × 1000 pixels resolution, obtained at 40× magnification, as well as 14 testing images at the same resolution. Dense annotations of all nuclei in all images are provided.

**CoNSeP:** This dataset Graham et al. [2019] contains 27 training images at 1000 × 1000 pixels resolution, obtained at 40× magnification, as well as 14 similar testing images. There are dense annotations of all nuclei in all images.

### 3.2 Experimental Set-up

**Baseline Methods:** The simplest supervised baseline method is to obtain the optimal intensity threshold using the training images. The images are converted to grey-scale and an optimal threshold to obtain binary segmentation is obtained. Additionally, a CNN of the same architecture as the confidence network, \(f_b\), is used as a supervised baseline. And finally, these results are also compared to the self-supervised method that uses scale prediction as a pretext [Sahasrabudhe et al. 2020], which is referred to as the Scale PrecText method. All baselines are evaluated on both mentioned datasets.

<table>
<thead>
<tr>
<th># Patches</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dice</td>
<td>0.6643</td>
<td>0.7071</td>
<td>0.6682</td>
<td>0.6986</td>
<td>0.6059</td>
</tr>
<tr>
<td>Patch Size</td>
<td>20²</td>
<td>30²</td>
<td>50²</td>
<td>80²</td>
<td>120²</td>
</tr>
<tr>
<td>Dice</td>
<td>0.5063</td>
<td>0.6562</td>
<td>0.6980</td>
<td>0.6606</td>
<td>0.6871</td>
</tr>
</tbody>
</table>

Table 1: The median validation Dice score over three runs on MoNuSeg is used to select number of patches and patch sizes (shown in boldface).

**Model Hyperparameters:** The default configuration uses a confidence network, \(f_b\), closely inspired by Sahasrabudhe et al. [2020], and described in Appendix 2.3. Entropy based stochastic sampling described in Section 2.3 and \(K = 4\) classes are used. The patch size of 50 × 50 pixels and 10 patches from each image.

---

2MoNuSeg data is available at [https://monuseg.grand-challenge.org/Data/](https://monuseg.grand-challenge.org/Data/)

3CoNSeP data is available at [https://warwick.ac.uk/fac/cross_fac/tia/data/hovernet/](https://warwick.ac.uk/fac/cross_fac/tia/data/hovernet/)
in a batch are chosen based on experiments on the MoNuSeg dataset, as reported in Table 1. The proposed framework was trained for a maximum of 300 epochs. To limit RAM usage, the input images are cropped into 300 × 300 pixels, where the location is selected uniformly at random as to add some data variance. No other pre-processing nor data augmentation is applied. Batch size of 10 is used for all experiments.

**Implementation:** The proposed framework is implemented in PyTorch [Paszke et al. 2019] using PyTorch Lightning [Falcon and team 2019], with support to be trained on GPUs; all training has been performed on a system with an NVIDIA GeForce RTX 3060 and intel-i7 processor with 32GB memory.

**Experiments:** With the objective of comparing the segmentation performance of the proposed CSL framework with the baselines, we perform experiments on each of the datasets described in Section 3.1. Our CSL framework is initialised with the hyper-parameters described in Section 3.2 on each of the datasets and the weights of the confidence network are randomly initialised. A single training epoch consists of contrasting 10 patches of size 50 × 50 (see Table 1) sampled from one random crop from each training image. At the end of each epoch, the confidence maps obtained from the confidence network are thresholded at $p = 0.5$ and the segmentation performance is evaluated using the labels for the training images by computing the Dice score. After training for 300 epochs, the model with highest validation Dice score is selected for evaluation.

---

### Table 2: Dice scores on test sets of MoNuSeg and CoNSeP datasets for the different methods along with maximum, mean+sd and 80th percentile scores for our method. Optimal intensity threshold method (Int. Thresh.), supervised CNN baseline, self-supervised method using scale prediction pretext task (Scale Pretext) and the proposed CSL framework. The number of trainable parameters and convergence time per method are also reported.

<table>
<thead>
<tr>
<th>Methods</th>
<th>MoNuSeg</th>
<th>CoNSeP</th>
<th># Par.</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimal Thresh.</td>
<td>0.5823</td>
<td>0.5700</td>
<td>1</td>
<td>&lt;1m</td>
</tr>
<tr>
<td>CNN</td>
<td>0.7666</td>
<td>0.7375</td>
<td>10.8k</td>
<td>≈10m</td>
</tr>
<tr>
<td>Scale Pretext</td>
<td>0.6209</td>
<td>0.5139</td>
<td>21.7M</td>
<td>&gt;12h</td>
</tr>
<tr>
<td><strong>Ours:</strong> Max.</td>
<td>0.6797</td>
<td>0.6266</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean ± s.d</td>
<td>0.62±0.05</td>
<td>0.52±0.06</td>
<td>10.8k</td>
<td>≈5m</td>
</tr>
<tr>
<td>80 perc.</td>
<td>0.6469</td>
<td>0.5448</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

4 Most likely this is a large under-estimation of the actual convergence time based on in-house implementation as the actual run-time was not reported in Sahasrabudhe et al. [2020].
MoNuSeg: Our method outperforms the supervised optimal threshold method when trained on the MoNuSeg dataset using MSE-similarity. Qualitative results on four randomly selected test set images from the MoNuSeg dataset for the model trained on WSI data are shown in Figure 3. Additional segmentation results are also shown in Appendix C.

CoNSeP: A similar performance trend is observed on the CoNSeP dataset, where we notice the mean Dice score of our method with the CE similarity measure performs better than the Scale Pretext method. Across all the experiments over all the datasets, the number of parameters of our method (10.8k) is several orders of magnitude fewer than Scale Pretext method (21M), and our method takes only a fraction of time until convergence (5m) compared to theirs (>12h). This fast convergence time has implications in overcoming the strong influence of the confidence maps obtained at initialisation (epoch-0) which can result in convergence to poor solutions. While some runs result in useless performance, dragging the mean down, good runs do happen frequently. We see that the 80’th percentile performance over multiple runs is often significantly better than the mean performance; thus running the model just a handful of times (which is still fast) will likely yield a well-performing one. As a standard practice, we therefore recommend running multiple repeats of our model on any dataset.

5 Discussions & Conclusions
Confidence maps to Segmentation masks: The efficient CSL framework presented in this work outputs multiple confidence maps corresponding to objects believed to belong to the same class. Going from these confidence maps to the segmentation masks by definition requires expert input. In this work, we alleviate this by using the labels for model selection/validation. For other unlabelled datasets, some other form of validation would be required to align the concept of a biomedical image foreground to that of a confidence map of a self-supervised framework. The segmentation performance based on the confidence maps obtained in different output channels is illustrated and discussed in Appendix E.2.

Self-supervised segmentation performance: The main baseline we compared our method to is the self-supervised Scale Pretext method [Sahasrabudhe et al. 2020]. The results of this method is not significantly different from ours. This is even with specific pre-processing (stain-normalisation) and elaborate post-processing followed in Sahasrabudhe et al. [2020], which are altogether left out in our work. Further, compared to the Scale Pretext method ours is significantly simpler, both in terms of the model complexity, the elaborate regularisation of their objective function and training time (5m versus 12h).

Predictions in different channels: Having multiple output classes increases the likelihood of the desired class to actually be segmented, at the cost of some computational resources. However, even with $K=4$ classes, there are runs where none of the classes correspond particularly well with the desired nuclei segmentation. Hence, a few of the runs will turn out useless, requiring training the framework multiple times, and choosing a good instance based on validation performance. Training our method multiple times is recommended, as it could benefit from the variance in segmentation performance. This is captured in Table 3 where the best score is often significantly better than the mean score.

Limitations: The main limitation in the proposed framework coincides with its simplicity — there are no constraints encouraging the model to actually segment the desired objects of interest, resulting only in small performance improvements. Further, obtaining segmentation masks of interest in one of the confidence maps is not guaranteed; so as a standard practice we suggest running the model multiple times. This might not end up being a major limitation, after all, as many model runs do converge to reasonable solutions as captured by the 80-percentile scores in Table 2. Finally, we assume that the similarity measure plays an important role in model performance. We experimented with a few of these measures. However, thorough investigations of more refined similarity measures and their influence on multi-class segmentation remains as future work.

Conclusions: We presented a self-supervised framework for segmenting nuclei from histopathology data, which uses patch-based contrastive learning. We introduced a novel technique to mine for positive- and negative- patches for contrasting based on the average entropy of the confidence maps. The segmentation performance based on the confidence maps obtained in different output channels is illustrated and discussed in Appendix E.2.
References


Appendix

Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Related Work</td>
<td>9</td>
</tr>
<tr>
<td>B</td>
<td>Sampling Methods</td>
<td>10</td>
</tr>
<tr>
<td>B.1</td>
<td>Confidence-based Sampling</td>
<td>10</td>
</tr>
<tr>
<td>B.2</td>
<td>Top-k sampling</td>
<td>10</td>
</tr>
<tr>
<td>C</td>
<td>Similarity Measures</td>
<td>10</td>
</tr>
<tr>
<td>D</td>
<td>Training Evolution</td>
<td>11</td>
</tr>
<tr>
<td>E</td>
<td>Model Hyperparameters and Training</td>
<td>11</td>
</tr>
<tr>
<td>E.1</td>
<td>Patch shape</td>
<td>11</td>
</tr>
<tr>
<td>E.2</td>
<td>Influence of Number of Output Channels</td>
<td>12</td>
</tr>
<tr>
<td>E.3</td>
<td>Confidence Network Architecture</td>
<td>12</td>
</tr>
<tr>
<td>F</td>
<td>Datasets visualisation</td>
<td>13</td>
</tr>
<tr>
<td>G</td>
<td>Additional Results</td>
<td>13</td>
</tr>
<tr>
<td>G.1</td>
<td>Qualitative Results</td>
<td>13</td>
</tr>
<tr>
<td>G.2</td>
<td>Confusion matrices</td>
<td>13</td>
</tr>
<tr>
<td>H</td>
<td>Experiments with Negative Results</td>
<td>13</td>
</tr>
<tr>
<td>H.1</td>
<td>TNBC dataset</td>
<td>13</td>
</tr>
<tr>
<td>H.2</td>
<td>Whole Slide Images in MoNuSeg</td>
<td>13</td>
</tr>
</tbody>
</table>

A Related Work

The high level ideas of contrastive learning and self-supervision have been widely studied in literature in recent years. Our work integrates some of these ideas into an efficient contrastive self-supervision framework that is able to handle complex biomedical imaging data.

Many CSL methods are trained on large unlabelled datasets using an assortment of pretext tasks. These tedious training processes are time consuming, sometimes requiring several days of compute time. This could have a net negative impact due to the increase in their energy consumption compared to supervised learning, which might converge to useful solutions faster.

Due to the availability of several well-curated natural image datasets, bulk of the CSL methods are trained on these data. The pre-training gains observed with self-supervised models on these natural image datasets do not immediately translate to more complex images that might not share features with natural images, such as the ones encountered in biomedical imaging.

Pretext task based self-supervision: Self-supervised learning is a form of unsupervised learning which uses proxy or pretext tasks to obtain useful representations of the data. These pretext tasks can be simple data augmentation strategies such as rotation and colourisation. However, more complex pretext tasks such as patch context prediction or solving image jigsaw puzzles have also shown to learn useful representations.

Pre-training with self-supervised learning: Most pretext based self-supervision methods are used as a pre-training step. These pre-trained models are further fine tuned, usually, using fewer labelled data than in fully supervised settings to achieve performance that is comparable to supervised methods, for instance to perform label efficient segmentation.

Contrastive self-supervised learning: Similarity based contrastive learning for obtaining discriminative representations has been studied in earlier works, but was formalised for self-supervised learning in the SimCLR framework. The key challenge of contrastive self-supervision, however, is mining for hard examples. Simple data augmentation strategies such as the ones used in SimCLR are not found to be very useful for discriminating complex features.

Self-supervision and biomedical image segmentation: Pretext tasks for improving nuclei segmentation from histopathology images have been formulated based on predicting the scaling factor of image regions. This is based on the assumption that the background class in tissue imaging usually appears similar at different scales compared to nuclei which vary between scales. By using a pretext task of
predicting the scale factor, the attention network is able to learn intermediate representations that are close to the nuclei segmentation mask. This scale based pretext task is very specific to the nuclei segmentation task, and might not generalise to other biomedical imaging data.

**B  Sampling Methods**

Several sampling procedures including the entropy based sampling presented in Section 2.3 have been tested during the development of this framework. We next present two most noteworthy alternatives as confidence based stochastic sampling and top-k sampling.

**B.1 Confidence-based Sampling**

This sampling method is based on the probability distribution of average patch confidences of the patches in $S$. The probability of selecting a patch $(R, i) \in S$ for the set of positive patches is

$$P(R) = \frac{A_k(R, i)}{\sum_{T \in S} A_k(T, j)}.$$  

A negative patch is sampled according to the probability distribution:

$$N(R) = \frac{1 - P(R)}{\sum_{T \in S} (1 - P(T))} = \frac{1 - P(R)}{|S| - \sum_{T \in S} P(T)} = \frac{1 - P(R)}{|S| - 1}.$$  

Notice that the sets of positive- and negative- patches are not necessarily disjoint with this sampling procedure. For simplicity, the sampling is performed with replacement – keeping all occurrences of selected patches – turning the sampled sets into multi-sets.

**B.2 Top-k sampling**

This sampling procedure partitions the set of candidate patches $S$ into positive- and negative- candidate sets, $P'_k$ and $N'_k$, respectively. Here, a patch $R$ is in $P'_k$ if it is among the $k$ patches with highest average patch confidence. Letting $k > n$ allows the introduction of some randomness by uniformly randomly sampling $n$ patches from $P'_k$ to the set of sampled positive patches and similarly for $N'_k$ into the set of sampled negative patches.

This sampling procedure performs reasonably well in our experiments for datasets with high contrast between features and does so in relatively few epochs. While the limited randomness leads to convergence in fewer epochs, it may be a limiting factor in harder datasets.

**C  Similarity Measures**

Given the confidence map scaled images, several choices for the similarity measure can be explored:

1. **Mean squared error (MSE)**: Given two patch locations, $(R_1, i)$ and $(R_2, j)$, with a bijective mapping, $g$, between their elements, the MSE similarity between them can be defined as

$$s((R_1, i), (R_2, j)) := \frac{1}{|R_1|} \sum_{u \in R_1} \left| F^k_i(u) - F^k_j(g(u)) \right|^2$$

where $i$ and $j$ may be the same image.

2. **Cross entropy (CE)**: Similar to the MSE similarity, a mean binary CE similarity between two patches $(R_1, i)$ and $(R_2, j)$ with a bijective map $g$ may be defined as

$$s((R_1, i), (R_2, j)) := \frac{1}{|R_1|} \sum_{u \in R_1} \left[ F^k_i(u) \log_2 F^k_j(g(u)) + (1 - F^k_i(u)) \log_2 (1 - F^k_j(g(u))) \right].$$

3. **Feature network**: Another approach to comparing patches, which illustrates the abstractness of the requirements for the similarity measure and possibilities for generalisation, is to introduce a featuriser, $F: \mathcal{X} \rightarrow \mathbb{R}^\ell$, which maps patches into an $\ell$-dimensional feature space, again based on $F$. This featuriser can e.g. be a convolutional neural network. The similarity measure can then be

$$s((R_1, i), (R_2, j)) := \frac{\langle F^k_i(R_1), F^k_j(R_2) \rangle}{\|F^k_i(R_1)\| \|F^k_j(R_2)\|}$$

using the cosine similarity, which is comparable to what is done in SimCLR [Chen et al. 2020].
Choice of similarity measures: The choice of similarity measure has an influence on our framework’s performance, and different choices may suit different datasets better. The similarity measures tested in this work are all very simple: either directly pixel based like MSE or CE or grey-scale density based like KL-divergence or mutual information. Neither of them perform well in all cases, and we speculate that this framework has a bigger potential with more refined similarity measures. In the end, we only use MSE and CE across datasets, and report the best performance achieved with either of the two.

D Training Evolution

Figure 1 illustrates how the trained model improves its delineation and segmentation performance, along with the sampled positive and negative patches. Figure 4 illustrates the distribution of the patch confidence as well as patch entropy, as a function of training epochs. As illustrated in that figure, the average patch confidence is initially distributed around 0.5, which corresponds to a lot of noise; the confidence network does not yet know what to select. This also corresponds to high entropies initially.

Then over time, a larger proportion of the patches gets either low or high confidence, and the patch entropy generally decreases. This increase of information — increased gap between low and high confidence — is the fundamental mechanism of the proposed framework. Driving patch confidence towards high or low is exactly the contrastive objective.

E Model Hyperparameters and Training

E.1 Patch shape

For efficient processing on GPUs, the patches are chosen to be equally sized, rectangular, and aligned to the pixel grid. Also requiring all input images \( i = 1, \ldots, M \) to be of the same size allows for further efficiency in implementation. The creation of the sets of candidate patches, \( W_k \), as well as the calculation of \( A_k(R,i) \) for all patches \( (R,i) \in W_k \) across channels \( k = 1, \ldots, K \), can be performed in a single convolutional operation; using a non-trainable convolution kernel of the patch size with
Figure 5: Validation Dice score during training, shown for all the $K=4$ output channels. The vertical line depicts the epoch of the best validation Dice score, and hence the version of the model which was later evaluated on the test set. Notice that channels 2 and 3 are almost following each other, indicating a redundant output channel.

Figure 6: Validation Dice score during training, shown for all the $K=4$ output channels. The vertical line depicts the epoch of the best validation Dice score, and hence the version of the model which was later evaluated on the test set. In this case, all four channels appear to be learning different features with channel 3 focusing on the foreground class.

all values being the reciprocal of the kernel size, utilising striding to adjust the number of patches to create. For further improved data-parallelism, this operation can also be scaled across all channels simultaneously by adding a dimension to the kernel’s weights and appropriately sparsifying the kernels with zero-entries.

E.2 Influence of Number of Output Channels

Increasing the number of output classes/Channels, $K$, increases the likelihood that at least one of the trained model’s channels actually segments the desired features — at least under the assumption that the different channels are trained towards recognising different features. The inter-channel contrastive loss is introduced to reward such class separation.

Figures 5 and 6 illustrate the validation Dice score computed during training of two models. As is evident from Figure 5, not all $K=4$ classes are properly separated. This can likely be explained by the intra-channel contrastive loss being a harsher penaliser than the inter-channel contrastive loss. This indicates that the inter-channel contrastive loss is not being enforced sufficiently.

Figure 6 illustrates predicted confidence maps and inferred segmentation maps from each of the $K=4$ channels in the model presented in Figure 5, at the epoch highlighted in that figure. Notice the same relation between channels as in Figure 5. Note that no morphological post-processing operations are performed, and that the segmentation performance is hence not representative of what is achieved in the final reporting. The model depicted here achieved a Dice score on the test set of 0.6797 after the post-processing.

E.3 Confidence Network Architecture

The confidence network, $f_\theta$, in all configurations of our model are implemented as a fully convolutional neural network with residual connections, summarised as a sequence of

Conv2dBlock, $s=1$, $p=d=1$

ResidualBlock
  Conv2dBlock, $s=1$, $p=d=1$
  Conv2dBlock, $s=1$, $p=d=1$

ResidualBlock
  Conv2dBlock, $s=1$, $p=d=2$
  Conv2dBlock, $s=1$, $p=d=2$

ResidualBlock
  Conv2dBlock, $s=1$, $p=d=3$
  Conv2dBlock, $s=1$, $p=d=3$

ResidualBlock
  Conv2dBlock, $s=1$, $p=d=5$
  Conv2dBlock, $s=1$, $p=d=5$

ResidualBlock
  Conv2dBlock, $s=1$, $p=d=10$
  Conv2dBlock, $s=1$, $p=d=10$

ResidualBlock
  Conv2dBlock, $s=1$, $p=d=20$
  Conv2dBlock, $s=1$, $p=d=20$

Sigmoid

where each Conv2dBlock consists of a sequence of a convolutional layer, batch normalisation, and ELU activation function, with $s$ denoting the stride, $p$ denoting the padding, and $d$ denoting the dilation of
the convolutional layer. All convolutional layers have kernel size $3 \times 3$. The first convolutional layer takes three input-channels (corresponding to colour images), and the last convolutional layer has $K$ output-channels. All other channel counts are 8. The ResidualBlocks computes the residual over the elements inside their indentation. This architecture is based on the attention network used in Sahasrabudhe et al. [2020].

Notice that, in the context of recent large-scale deep learning, the confidence network with its 10.8K trainable parameters is tiny. This has implications on the training time and the compute resources (such as maximum GPU memory required), making the overall method more efficient.

F Datasets visualisation

Figure 7 illustrates the variance of data in the three tested datasets. Three examples are randomly chosen from each dataset and presented along with their ground truth segmentation masks.

![Figure 7: Three pairs of images and corresponding ground truth segmentation masks (in green) from each of the three datasets: MoNuSeg Kumar et al. [2017] (top row); TNBC Naylor et al. [2018] (middle row); and CoNSeP Graham et al. [2019] (bottom row). Notice differences in nuclei (foreground) counts, sizes, and colours/intensities.](image)

G Additional Results

G.1 Qualitative Results

Figure 8 illustrates the prediction performance of a trained model using the proposed framework. This illustrates the necessity of multiple channels, as not all are useful for the intended task.

G.2 Confusion matrices

Table 3 shows confusion matrices of proposed framework and the optimal threshold, respectively, on the MoNuSeg Kumar et al. [2017] dataset. Notice that the numbers correspond to the foreground constituting a minority of the total images.

<table>
<thead>
<tr>
<th></th>
<th>Actual/Pred.</th>
<th>P</th>
<th>N</th>
<th>Actual/Pred.</th>
<th>P</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>0.101</td>
<td>0.115</td>
<td>P</td>
<td>0.145</td>
<td>0.072</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>0.333</td>
<td>0.452</td>
<td>N</td>
<td>0.082</td>
<td>0.703</td>
</tr>
</tbody>
</table>

Table 3: Confusion matrices for the optimal thresholding method and the proposed framework on MoNuSeg

H Experiments with Negative Results

H.1 TNBC dataset

In addition to the two histopathology datasets reported in the main paper, we also experimented on a third dataset: TNBC dataset Naylor et al. [2018]. This dataset consists of 50 images at $512 \times 512$ pixels, along with dense annotations of all nuclei. We split them into a training set of 28 images and testing set with the remaining 22 images; filenames ending 1 and 2 are used as test set, the remaining as train set.

On the TNBC dataset, both versions (MSE, CE) of our method under-perform, as shown in Table 4. This could be due to the fact that the nuclei of interest have regions inside of them which are similar to the background class, according to the two similarity measures used. Another reason could also be the extreme class imbalance between the foreground and background pixels in this dataset, as seen in Figure 7.

H.2 Whole Slide Images in MoNuSeg

The images in the MoNuSeg dataset originate from whole slide images (WSIs) available at the TCGA portal. Inspired by Sahasrabudhe et al. [2020] we fetch the 37 corresponding WSIs, from which the MoNuSeg training set originates, as a separate dataset.

5Whole slide images are available at [https://portal.gdc.cancer.gov/](https://portal.gdc.cancer.gov/)
Figure 8: Predictions on validation set during training. Same model as Figure 5 and at the epoch with the best validation Dice score, which is depicted by the vertical line in that figure. Top row: Input images from validation set. Second row: Ground truth segmentation masks. In each of the $K = 4$ remaining pairs of rows: Upper row: Confidence map from that channel; Lower row: Predicted segmentation obtained by thresholding the confidence map, colour-coded as in Figure 3 where blue is true positive regions, green indicates false negatives, and red indicates false positives. No post-processing operations are performed.

and use the 37 labelled MoNuSeg training images for validation of the segmentation performance and use 14 images as test set.

When trained on the additional WSI dataset, the Scale Pretext method appears to be benefit and outperforms our method 4. This indicates that additional training of our CSL method does not necessarily achieve performance gains. We speculate the simple FCNN (with 10.8k parameters) is unable to learn additional features from the new data. We hope to experiment with more complex confidence networks to test this hypothesis in future work.
<table>
<thead>
<tr>
<th>Methods</th>
<th>MoNuSeg Subset</th>
<th>MoNuSeg Full</th>
<th>CoNSeP</th>
<th>TNBC</th>
<th># Param.</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Int. Thresh.</td>
<td>0.5823</td>
<td>0.6205</td>
<td>0.5700</td>
<td>0.5590</td>
<td>1</td>
<td>&lt;1m</td>
</tr>
<tr>
<td>CNN</td>
<td>0.7666</td>
<td>0.7867</td>
<td>0.7375</td>
<td>0.7121</td>
<td>10.8k</td>
<td>≈ 10m</td>
</tr>
<tr>
<td>Scale Pretext</td>
<td>0.6209</td>
<td>–</td>
<td>0.7477</td>
<td>0.5139</td>
<td>21.7M</td>
<td>&gt; 12h</td>
</tr>
</tbody>
</table>

Ours (MSE): Max. 80 perc. Mean ± s.d
<table>
<thead>
<tr>
<th></th>
<th>Mean ± s.d.</th>
<th>Mean ± s.d.</th>
<th>Mean ± s.d.</th>
<th>Mean ± s.d.</th>
<th>Mean ± s.d.</th>
<th>Mean ± s.d.</th>
<th>Mean ± s.d.</th>
<th>Mean ± s.d.</th>
<th>Mean ± s.d.</th>
<th>Mean ± s.d.</th>
<th>Mean ± s.d.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Int. Thresh.</td>
<td>0.6797</td>
<td>0.7088</td>
<td>0.7181</td>
<td>0.5507</td>
<td>0.4324</td>
<td>0.59 ± 0.05</td>
<td>0.61 ± 0.17</td>
<td>0.62 ± 0.12</td>
<td>0.45 ± 0.08</td>
<td>0.32 ± 0.11</td>
<td>10.8k</td>
</tr>
<tr>
<td>CNN</td>
<td>0.6488</td>
<td>0.6176</td>
<td>0.6647</td>
<td>0.6266</td>
<td>0.5101</td>
<td>0.59 ± 0.05</td>
<td>0.52 ± 0.18</td>
<td>0.63 ± 0.02</td>
<td>0.52 ± 0.06</td>
<td>0.37 ± 0.20</td>
<td>10.8k</td>
</tr>
</tbody>
</table>

Table 4: Dice scores on test sets of the three datasets for the different methods (maximum, 80'th percentile, and mean±sd). Optimal intensity threshold method (Int. Thresh.), supervised CNN baseline, self-supervised method using scale prediction pretext task (Scale Pretext) and the proposed CSL framework using two similarity measures: mean-squared error (MSE) and cross-entropy (CE). Convergence time and the number of trainable parameters per method are also reported.