FocAnnot: Patch-Wise Active Learning for Intensive Cell Image Segmentation

Bo Lin[1], Shuiguang Deng[1(✉)], Jianwei Yin[1(✉)], Jindi Zhang[1], Ying Li[1], and Honghao Gao[2,3(✉)]

[1] College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China
{rainbowlin,dengsg,zjuyjw,zjindiss,cnliying}@zju.edu.cn
[2] School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China
gaohonghao@shu.edu.cn
[3] Gachon University, Seongnam, Gyeonggi-Do 461-701, South Korea

Abstract. In the era of deep learning, data annotation becomes an essential but costly work, especially for the biomedical image segmentation task. To tackle this problem, active learning (AL) aims to select and annotate a part of available images for modeling while retaining accurate segmentation. Existing AL methods usually treat an image as a whole during the selection. However, for an intensive cell image that includes similar cell objects, annotating all similar objects would bring duplication of efforts and have little benefit to the segmentation model. In this study, we present a patch-wise active learning method, namely FocAnnot (focal annotation), to avoid such worthless annotation. The main idea is to group different regions of images to discriminate duplicate content, then evaluate novel image patches by a proposed cluster-instance double ranking algorithm. Instead of the whole image, experts only need to annotate specific regions within an image. This reduces the annotation workload. Experiments on the real-world dataset demonstrate that FocAnnot can save about 15% annotation cost to obtain an accurate segmentation model or provide a 2% performance improvement at the same cost.

Keywords: Active learning · Intensive cell image · Duplicate annotation · Semantic segmentation

1 Introduction

Semantic segmentation is a fundamental and challenging task in computer vision. Given a single image, it aims to distinguish and localize each predetermined object at the pixel level. Owing to the rapid development of deep learning in recent years, advanced data-driven models such as fully convolutional network (FCN) [17] and Deeplab [4] can automatically discriminate multiple objects in an intricate image with promising results, which are faster and more accurate than old approaches. Many applications have introduced semantic segmentation techniques to enhance

H. Gao et al. (Eds.): CollaborateCom 2020, LNICST 350, pp. 355–371, 2021.
https://doi.org/10.1007/978-3-030-67540-0_21

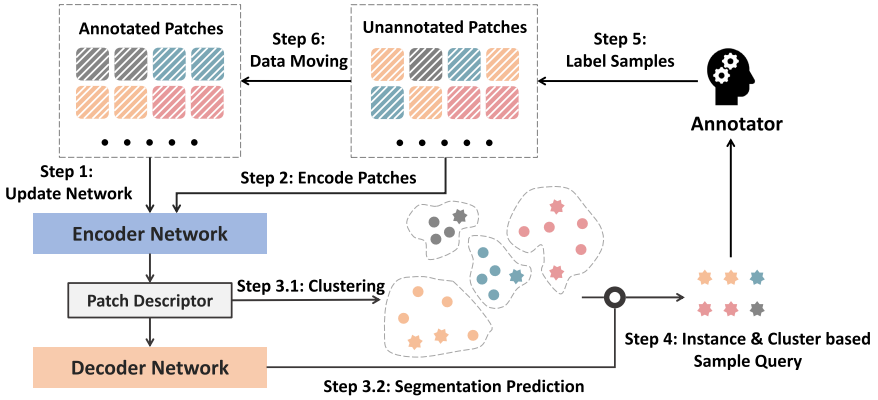


Fig. 1. Overview of our approach. All annotated and unannotated images are partitioned into patches before sending to the loop. Six steps are executed in order and repeated until reaching stop criteria.

automation level, such as remote sensing monitoring, autonomous vehicles, and auxiliary diagnosis [1,3,7,27]. Besides the superiority of the model structure and learning algorithm, the success of deep learning also relies on high-quality labeled data. Unfortunately, this data requirement cannot be satisfied in many practical problems, for instance, the biomedical image segmentation task. Annotating this kind of images is very costly and time-consuming because only pathologists are able to identify tissues or lesions, and mark their contours.

To this end, active learning (AL), which intends to maximize the model performance with minimum cost in labeling data, becomes an emerging research hotspot. In other words, an AL method iteratively runs a query strategy to select the most valuable samples for the annotation and helps classifiers achieve high accuracy from limited data points [2,14,22]. Many studies have introduced how to combine AL to the biomedical image segmentation task to reduce annotation cost [6,19,20,29]. In this work, we focus on the intensive cell image, within which cell objects are close together. We observe that cells in this kind of image have relatively fixed and similar contours. Thus, existing AL methods that run an image-level query strategy during the data selection would bring duplicate annotation because similar objects provide limited information in model training. This inconsistency raises an interesting problem: *Is the image-level query strategy efficient enough for the intensive cell image segmentation?* Based on the observation, we assume that further cost reduction can be achieved by measuring different regions of images separately and only selecting the most critical regions within images for the annotation.

Consequently, we present a patch-wise active learning method named FocAnnot (focal annotation) for the intensive cell image segmentation, as illustrated in Fig. 1. There are six steps in a selection iteration. In Step 1, annotated images are partitioned into patches with fixed size to initialize an encoder-decoder convolutional network (ED-ConvNet) for image segmentation. The unannotated

candidates with the same patch size are then fed to the trained ED-ConvNet to get segmentation results (Step 3.2), as well as their latent representations mapped by the encoder part of ED-ConvNet (Step 2). After that, a clustering method is adopted to cluster candidates into distinct groups (Step 3.1). In Step 4, we propose a novel query strategy that integrates candidate information at both instance and cluster level generated in Step 3. Based on query results, we select valuable regions within images to experts for the annotation (Step 5). Finally, these new samples are added to enlarge the annotated dataset in Step 6. The selection procedure is repeated until reaching predefined conditions.

The main contributions of this work can be summarized as follows:

- We propose FocAnnot, a patch-wise active learning method, to reduce duplicate annotation and save cost in the intensive cell image segmentation task. FocAnnot measures distinct regions within an image and only asks experts to annotate a part of valuable regions.
- We propose a cluster-instance double ranking query strategy consisting of two cluster-based criteria that estimate the importance of different image patch groups, and an instance criterion incorporated with traditional uncertainty for the patch selection.
- FocAnnot is evaluated on a real-world cell-intensive dataset. Up to 15% cost saving or 2% performance improvement can be achieved in the segmentation task.

The remainder of this paper is organized as follows. In Sect. 2, we review the related work on biomedical image segmentation and active learning. Section 3 describes the details of the proposed FocAnnot. Experimental results on a real-world dataset are reported in Sect. 4, and conclusions are given in Sect. 5.

2 Related Work

Active Learning. The main task of AL is to design a query strategy that measures the value of unlabeled data for different task objectives [6,10,14,16]. Recent studies can be concluded in three categories, i.e., single model, multi-model, and density-based methods. In the class of single model, the most informative samples are picked according to the probabilistic outputs of a trained classifier. This strategy is also known as uncertainty sampling [8,30]. Similarly, the multi-model approach, or query-by-committee, also leverages predicted labels but in an ensemble way. A sample with the most disagreements among multiple classifiers is considered as an informative instance [12]. Kullback–Leibler (KL) divergence, entropy, and top-k best are some criteria in common use that measure amount of additional information brought by selected data. The density-based method aims to find data points that are uncertain as well as representative. The idea is re-ranking queried samples based on the similarity to their neighbors or directly querying from the pre-clustered sets [23]. Moreover, Zhou *et al.* [34] use a pre-trained CNN to predict augmented data based on the assumption that images

generated from the same seed are expected to have similar predictions. Entropy and KL divergence are employed to evaluate uncertainty and prediction consistency among augmented images, respectively. The study in [18] incorporates the generative adversarial network into the AL framework to generate informative data, while [15] introduces a learnable query strategy that estimates expected error reduction by a regressor. Yang *et al.* [29] applies uncertainty sampling to choose several candidates and discards duplicate selections with high similarity. Their proposed suggestive annotation method achieves state-of-the-art segmentation performance in an intensive cell image dataset.

Biomedical Image Segmentation. Automatic segmentation brings benefits to the medical field that enhances lesion identification, surgery planning, and evaluation of treatment effects. Recent advances in biomedical image segmentation have covered many organs, such as liver [11], brain [21], and prostate [33]. There are also many efforts on other human tissues, including cells [26], nucleus [28], and melanoma [32]. Technically, most of the state-of-the-art segmentation models are based on the convolutional neural network (CNN) with an encoder-decoder architecture. The encoder network applies convolution and down-sampling operation to images, which compresses raw inputs to learn latent features. The decoder network then deconvolves and up-samples latent features to predict each pixel in images. Many studies have designed new components to improve the robustness and generalization of CNN models. Ronneberger *et al.* [24] build skip connections between the down-sampling and up-sampling path to enhance the sharing of local information. Dilated convolution [31] is adopted to increase the receptive field, which works as the alternative to pooling operation but reduces the model size. Jointly learning knowledge from multi-scale data is an effective strategy as well, for instance, multiresolution inputs [5], multi-scale latent representations [4], and sequential structure of multiple networks [9].

3 Our Approach

Well-annotated data empowers segmentation models to achieve promising results but is also costly, especially for the biomedical image. With the constraint of annotation costs, active learning aims to retrieve the most valuable images from the unannotated dataset for the specific tasks. We suppose annotators are experts who provide high-quality labeling. Hence our task becomes to get similar segmentation performance of full supervision by selecting limited data with minimum cost. To this end, we propose FocAnnot, a patch-wise active learning method that only assigns parts of the region within an image to human annotators, to further reduce annotation costs of intensive cell images compared to existing image-level AL methods.

3.1 Overview of FocAnnot

The overview of our approach is shown in Fig. 1, including six steps. Images are first partitioned into small patches before sending them to the loop. The details

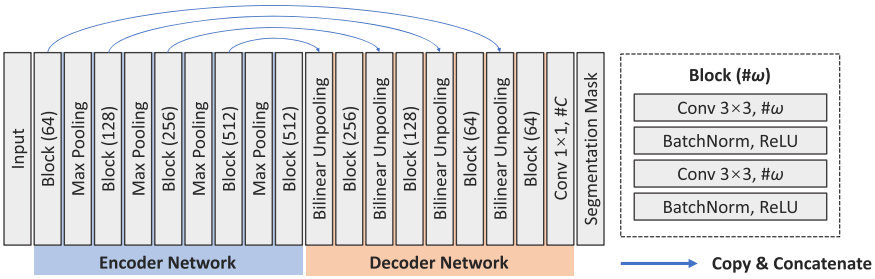


Fig. 2. Architecture of the ED-ConvNet implemented in this work. The symbols C and ω denote the number of classes and the number of kernels, respectively.

will be described in the next section. Initially, annotated image patches are fed to ED-ConvNet for network training. The encoder of ED-ConvNet learns latent representations of patches in low dimensional space and the decoding part makes predictions to each pixel. After that, we put all unannotated patches into well-trained ED-ConvNet. Besides decoding results, outputs of the encoder network are taken as well (*patch descriptors* for short). As mentioned earlier, cells within an image are similar. Instead of the overall profile, we focus on the local difference and group high-level features of patches, i.e., patch descriptors, into clusters in the third step. Each cluster can be regarded as regions with similar types of contents. Based on this, we propose a double ranking query strategy from perspectives of the image patch itself (instance-level) and its kind (cluster-level). Patches with the highest-ranking are selected for labeling and moved to the annotated dataset in the last two steps. Again, parameters of ED-ConvNet are updated by the enlarged training set. FocAnnot runs all these steps sequentially until predefined conditions are achieved, for instance, budget or accuracy of segmentation.

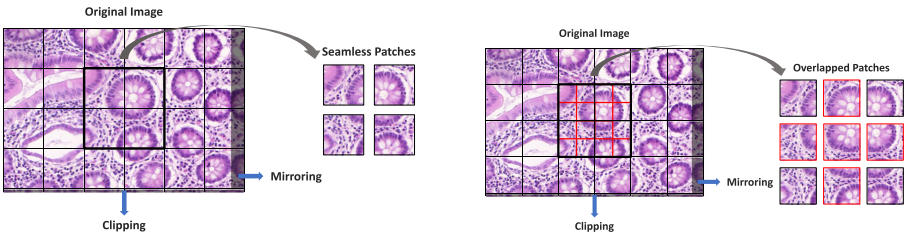
Furthermore, we illustrate an implementation of the ED-ConvNet constructed in FocAnnot. As shown in Fig. 2, ED-ConvNet contains a simple block design that executes a 3×3 convolution followed by a batch normalization [13] and ReLU activation, in two iterations. In the encoding part, a max-pooling layer is set between each of two blocks. An image patch is processed in such stacked layers and its patch descriptor is obtained, that is, the outputs of the second "Block (512)" in this example. The internal outputs of blocks are copied to the decoder network as one of the inputs in corresponding blocks. The decoding part has a symmetrical structure to the encoder network that uses bilinear interpolation for up-sampling to reconstruct image patches. At last, a 1×1 convolution layer is applied to predict the class of each pixel and gives final segmentation results.

3.2 Image Partitioning and Clustering

This preprocessing step aims to alleviate the problem of duplicate annotation in existing image-level AL methods. For cell-intensive images, the same types of biomedical objects are similar, with partial differences. Hence we can partition images into patches to focus on details at the region level. The advantage of

patch-wise learning is that each patch is considered to be of different importance. Compared to the image-wise annotation, experts are only required to annotate the most valuable regions in an image and ignore other parts, which save expenses.

We assume patch size and step size are two critical factors affecting outcomes, i.e., trade-offs between cost and accuracy. Patch size controls the integrity of partitioned objects. It should be neither too big nor too small, which goes against the idea of patch-wise learning or loses object information, respectively. Step size aims to increase the richness and novelty of partitioned objects. Similarly, small step size is not desirable because it increases the computational efforts of the query.



(a) Seamless partition ($l = 128$, $\rho = 0$) (b) Overlapped partition ($l = 128$, $\rho = 0.5$)

Fig. 3. Two strategies for image patches partition. Each patch is square with the side length of l, and ρ is ratio of overlapping between neighbors.

In this study, we set two partition strategies for cell-intensive images, as shown in Fig. 3. Patch size l and overlapping rate ρ are two parameters to control the partitioning. The patch size denotes the side length of each rectangular region while ρ indicates how many overlaps are between two adjacent patches. Figure 3a illustrates a seamless strategy that outputs 24 patches of 128×128 pixels. We zoom in a subregion involving four patches (bold black box) and show them on the right side of the image. Complete objects are scattered in different patches, but they are almost distinguishable from the background. In other words, the appropriate patch size has a limited impact on task difficulty. Figure 3b shows the overlapped partition strategy with $\rho = 0.5$. We add red lines to the same subregion to demonstrate the boundaries of additional patches. On the right side, extra five patches are obtained by half overlapping, which enriches the novelty of the unannotated dataset. It also helps alleviate the potential object integrity problem brought by the seamless strategy. Nevertheless, the side effect is the time of the sample query in Step 4 will increase because of the doubled data. In both two strategies, remaining areas on the bottom and rightmost of an image (gray shadows) will be dropped or expanded. An image is padded to fill the missing by mirroring regions if the required size is smaller than $l/2$, otherwise just clipping the area for alignment.

After partitioning, we are able to measure different regions separately to avoid duplicate annotation by the image-level selection. Various types of patches

are categorized into groups based on their similarity by calculating Euclidean distance between their patch descriptors and group centers:

$$\|\overline{\mathbf{F}} - \mathbf{M}\|^2 \tag{1}$$

where $\overline{\mathbf{F}}$ and $\mathbf{M}$ are channel-wise average pooling of a patch descriptor and the centroid of a group, respectively. In the remainder of this article, we use $\overline{\mathbf{F}}$ to denote the patch descriptor.

To improve query efficiency, we adopt the mini-batch k-means algorithm for the clustering. The number of groups depends on the complexity of cell objects. We have mentioned that the contours of some cells are quite simple. So a small k is enough to distinguish different types of patches for most cases. Detailed performance comparisons of the group number will be discussed in the experiment.

3.3 Cluster-Instance Double Ranking

Many criteria have been studied to determine which images are valuable for the annotation. Among them, uncertainty and diversity are two concepts to evaluate so-called "worthiness". The term uncertainty indicates how confident predictions are given by a model, while diversity considers the degree of dissimilarity among samples. Most of the current works focus on query strategy at the instance level but ignore the cluster structure. In this study, we propose two cluster-wise criteria that measure patch types based on uncertainty and diversity. At the instance level, we also propose a Wasserstein distance-based diversity criterion incorporated with an instance uncertainty criterion.

For an image patch $\mathbf{x}_n$, $p_j(\mathbf{x}_n)$ denotes the predicted probability of each pixel belonging to the j-th object by the trained ED-ConvNet. The segmentation results are close to 0.5 if the model does not have enough knowledge, for example, representative training data, to identify a patch. This kind of patch is informative and can be regarded as a candidate for the annotation. It is a good choice to use entropy to capture the degree of information involved in samples. The higher value suggests more uncertainty of a patch to the model. Thus, we define *instance uncertainty* (IU) as:

$$\mathcal{H}_n = -\sum_{j=1}^{C} p_j(\mathbf{x}_n) \log(p_j(\mathbf{x}_n)) \tag{2}$$

where C is the number of predefined objects.

Besides, we also want to select a couple of samples, among which are dissimilar. Two patches with high uncertainty could also cause duplicate annotation if they provide similar information. In our case, a patch descriptor can be seen as a probability distribution of high-level features. Thus, we introduce Wasserstein distance as a diversity measurement that estimates differences between two probability distribution P and Q. Compared to KL divergence, Wasserstein distance is a symmetric metric and satisfies the triangle inequality, which is suitable for the similarity calculation. Another advantage of Wasserstein distance is the

ability to measure two distributions with little overlap. For example, two patch descriptors would show quite different distribution even in the same group with similar IU. Jensen–Shannon divergence [34] is hard to measure the diversity of patches in this situation, while Wasserstein distance provides a better estimation. Let Ω be a metric space with distance function D and collection $\mathrm{Z}(P, Q)$ denotes all possible joint distributions on $\Omega \times \Omega$. The τ-Wasserstein distance is formalized:

$$\mathcal{W}_\tau(P, Q) = \left(\inf_{\zeta \in \mathrm{Z}(P,Q)} \int_{\Omega \times \Omega} D(u, v)^\tau \, d\zeta(u, v) \right)^{1/\tau} \tag{3}$$

Here ζ is a joint distribution of P and Q, and (u, v) is a data point sampled from ζ. Specially, we calculate Wasserstein distance between patch descriptors to measure the diversity. Supposing $\overline{\mathbf{F}}_n$ and $\overline{\mathbf{F}}_{n'}$ are two d-dimensional descriptors of image patches $\mathbf{x}_n$ and $\mathbf{x}_{n'}$, the *instance diversity* (ID) is defined as their 2-Wasserstein distance:

$$\mathcal{W}_{n,n'} = \min_{\mathcal{V} \in \mathfrak{G}(d)} \left(\sum_{m=1}^{d} \| \overline{\mathbf{F}}_{n,m} - \overline{\mathbf{F}}_{n',\mathcal{V}_m} \|^2 \right) \tag{4}$$

where $\mathfrak{G}(d)$ is all permutations of indices $\{1, \ldots, d\}$, and $\mathcal{V}$ is one of the permutations. The Euclidean norm is adopted as the distance function.

With IU and ID, each image patch can be scored and ranked. Only the first few candidates with the highest scores are selected. To get better results, the reweighting technique is applied for rank adjustment during the selection. For an image patch $\mathbf{x}_n$, its score is defined as the uncertainty of model predictions weighted by average diversity to other Q candidates:

$$\mathcal{S}_n = \mathcal{H}_n \times \frac{1}{Q} \sum_{q=1}^{Q} \mathcal{W}_{n,q} \tag{5}$$

Besides criteria at the instance level, we further describe two measurements that consider the characteristics of clusters. After applying the mini-batch k-means algorithm, patches with similar contents are grouped to the same cluster. We suppose that the degree of aggregation, or density, is relevant to the amount of information involved in a cluster. A dense cluster is less-informative because patch descriptors in corresponding metric space tend to be compact. On the contrary, a large distance between patch descriptors implies more novelty and uncertainty in a sparse cluster. To estimate informativeness of a cluster, we denote *cluster uncertainty* (CU) by the average distance of a single group:

$$\mathcal{I}_a^{(c)} = \frac{1}{|\mathcal{K}^{(c)}|} \sum_{\overline{\mathbf{F}}_n \in \mathcal{K}^{(c)}} \| \overline{\mathbf{F}}_n - \mathbf{M}_c \|^2 \tag{6}$$

where $\overline{\mathbf{F}}_n$ is a patch descriptor in the cluster $\mathcal{K}^{(c)}$, and $\mathbf{M}_c$ is the cluster centroid. The $|\cdot|$ indicates the size of a set.

Similar to the instance diversity, clusters should be dissimilar as well. Directly excluding limited clusters is infeasible. Alternatively, we determine the number of candidates to be provided in each group. A cluster that is far from others is asked for more patches because it is quite different from other groups and has a higher probability of providing valuable data. For a cluster $\mathcal{K}^{(c)}$, the mean distance to other centroids are defined as the *cluster diversity* (CD):

$$\mathcal{I}_r^{(c)} = \frac{1}{k-1} \sum_{v=1}^{k} \|\mathbf{M}_c - \mathbf{M}_v\|^2 \tag{7}$$

Instead of uniform sampling, we select the most valuable image patches from clusters proportionately based on the importance. Two cluster-wise criteria defined on Eq. (6) and Eq. (7) are parameterized by λ to give an estimation of cluster importance:

$$\mathcal{I}^{(c)} = \lambda \mathcal{I}_a^{(c)} + (1-\lambda)\mathcal{I}_r^{(c)} \tag{8}$$

At last, the proposed cluster-instance double ranking query strategy is described in Algorithm 1. The importance $\mathcal{I}^{(c)}$ of each cluster is calculated and normalized. Recall that $\sum_{c=1}^{k} \mathcal{I}^{(c)} = 1$, which means $\mathcal{I}^{(c)}$ can be treated as the probability of a cluster to provide informative data. For the double ranking strategy, we use $\mathcal{I}^{(c)}$ to confirm the number of required patches in each ranking step. In the first round, we filter out less informative data and retain $Q^{(c)}$ patches from cluster $\mathcal{K}^{(c)}$ based on IU. Corresponding reweighted scores $\mathcal{S}^{(c')}$ of only $Q^{(c)}$ patches are then calculated in the second step. Finally, we pick top $T^{(c)}$ rankings from refined cluster $\mathcal{K}^{(c')}$ for the annotation.

Algorithm 1
Cluster-instance double ranking query strategy

Input:
Patch descriptors in k groups $\mathcal{K}^{(1)}, \ldots, \mathcal{K}^{(k)}$;
The number of uncertain patches Q;
The number of required patches T in a query

Output:
Set of selected patches $\mathcal{A}$ for the annotation

$\mathcal{A} = \varnothing$
for $c = 1, \ldots, k$ **do**
  Compute $\mathcal{I}^{(c)}$ of $\mathcal{K}^{(c)}$ by Eq. (6), (7), (8)
  $Normalize(\mathcal{I}^{(c)})$
  $T^{(c)} = \mathcal{I}^{(c)} \times T$
  $Q^{(c)} = \mathcal{I}^{(c)} \times Q$
  Let $\mathbf{X}^{(c)} = \{\mathbf{x}_n | \overline{\mathbf{F}}_n \in \mathcal{K}^{(c)}\}$ where $\mathbf{x}_n$ is the image patch of $\overline{\mathbf{F}}_n$
  $\mathcal{H}^{(c)} = \{\mathcal{H}_n | \mathbf{x}_n \in \mathbf{X}^{(c)}\}$ by Eq. (2)
  $\mathcal{K}^{(c')} = TopRank(\mathcal{K}^{(c)}, \mathcal{H}^{(c)}, Q^{(c)})$
  $\mathcal{S}^{(c')} = \{\mathcal{S}_n | \mathcal{H}^{(c')}, \overline{\mathbf{F}}_n \in \mathcal{K}^{(c')}\}$ by Eq. (5)
  $\mathcal{A} = \mathcal{A} \cup TopRank(\mathcal{K}^{(c')}, \mathcal{S}^{(c')}, T^{(c)})$
end for
return $\mathcal{A}$

4 Experiment

We implemented our FocAnnot in Python using deep learning framework PyTorch and machine learning framework Scikit-learn. All experiments run on an Ubuntu server with eight cores of 2.20 GHz Intel Xeon E5-2630 and two NVIDIA GTX 1080 Ti GPU.

4.1 Dataset

The proposed active learning method is evaluated on a real-world cell-intensive dataset provided by the 2015 MICCAI Gland Segmentation Challenge (GlaS) [25], which contains 165 images of colon histology. As an example, Fig. 3 shows several glands involved in an image. According to the rules of GlaS, images are divided into a training set with 85 images and two test sets with 80 images in total (60 in Part A and 20 in Part B, P-A and P-B for short). In order to eliminate influence by such man-made split, we also generate a random train-test split from 165 images (Mixed for short) using 80% images as the training set and the remaining 20% for testing. All experiments will run on the three different train-test pairs.

Table 1. List of seven query strategies for comparison.

#Strategy	Partitioning	Clustering	Criteria
1	✗	✗	IU
2	✗	✗	SA
3	✗	✗	ED-ConvNet-SA
4	✓	✗	IU
5	✓	✗	IU+ID
6	✓	✓	IU+ID
7	✓	✓	IU+ID+CU+CD

4.2 Experimental Settings

In Algorithm 1, we set $Q = 5\%$ and $T = 2.5\%$ for a single selection, that is, FocAnnot queries 5% patches as candidates once from the training set and half of them are finally selected for the annotation. In each round, ED-ConvNet is retrained on the newly updated dataset, and then the segmentation performance is evaluated on the test set without image partitioning. In the experiment, the annotation cost is calculated as the number of pixels in the selected images. The parameter λ of Eq. 8 is fixed to 0.5. We repeat the train-test step until half of the training set is selected. Furthermore, we intend to explore the influence of parameter settings on the model performance, such as patch size, overlapping

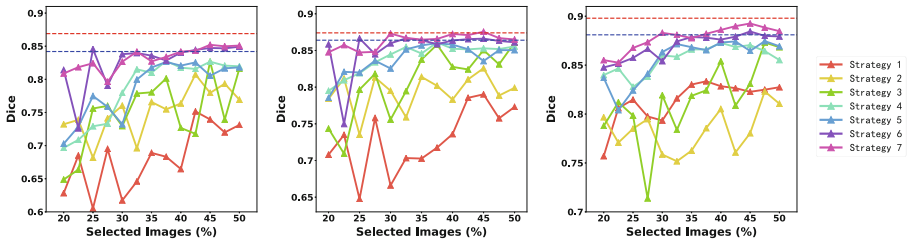
rate, and the number of clusters. Totally 12 combinations of $l = \{64, 128, 256\}$, $\rho = \{0, 0.5\}$, $c = \{3, 5\}$ are investigated on GlaS dataset. Two metrics, i.e., Dice coefficient and volumetric overlap error (VOE), are used for performance evaluation. Among them, Dice is preferred by the biomedical image segmentation task, so we choose it as the primary metric in comparisons. Formally, Dice and VOE are defined as:

$$Dice = \frac{2 \times |\mathbf{y} \cap \mathbf{y}'|}{|\mathbf{y}| + |\mathbf{y}'|} \tag{9}$$

$$VOE = 1 - \frac{|\mathbf{y} \cap \mathbf{y}'|}{|\mathbf{y} \cup \mathbf{y}'|} \times 100\% \tag{10}$$

Here $\mathbf{y}$ and $\mathbf{y}'$ are the ground truth and predicted segmentation result, respectively. The $|\cdot|$ is the number of pixels involved in this image. The larger Dice value or smaller VOE value indicates better performance of the model.

In Table 1, we list seven query strategies, including the state-of-the-art method in this scenario (suggestive annotation [29], SA for short), its variant ED-ConvNet-SA (replacing FCN in SA by our implemented ED-ConvNet) and our proposed criteria. Strategies 1–3 are baselines using the whole image while strategies 4–7 run a patch-wise selection. Specially, the comparison of strategies 1 and 4 investigate the effectiveness of image partitioning based on the traditional IU criterion. For strategies 2 and 3, we don not set contrast tests because SA and ED-ConvNet-SA both require a similarity computation with $\frac{1}{2}\mathcal{O}(N^2)$ time complexity, which is unacceptable on partitioned patches. In strategy 5, the proposed instance diversity is incorporated with instance uncertainty as a new selection criterion. In addition to the instance-level strategy, patches are selected from clusters equally in strategy 6, and proportionately in strategy 7. Note that query strategy 7 is a full implementation of cluster-instance double ranking strategy described in Algorithm 1.



(a) P-A (b) P-B (c) Mixed

Fig. 4. Comparisons between seven query strategies evaluated on three train-test pairs. Red and blue dash lines are segmentation performance of ED-ConvNet and FCN module in SA using full training data, respectively.

4.3 Results

We run seven query strategies listed in Table 1 using the same experimental settings. In strategies 4–7, the side length l, overlapping rate ρ, and group number k are set to 128, 0.5, and 3, respectively. For results in Fig. 4, the annotation cost of full-size images and image patches is aligned to the same scale based on the number of revealed pixels. It needs to be emphasized that the number of training images in Mixed is different from P-A and P-B, so the percentage of selected data cannot be compared directly. In the following, we first summarize our observations.

Strategy 1 vs. 2 vs. 3: Performance of Baselines. As we can see in Fig. 4a and Fig. 4b, SA indeed improves the segmentation performance compared to traditional IU. Surprisingly, in Fig. 4c, SA fails on mixed data and even perform worse than IU criterion. We believe the man-made interference in the provided train-test split leads to over-fitting and reduces the generalization of SA. In strategy 3, we replace all FCNs in SA by our ED-ConvNet to get its variant, i.e., ED-ConvNet-SA. Again, ED-ConvNet-SA cannot surpass IU on Mixed with less than 40% data. This also explains the limited generalization of the SA-based strategy. From a model point of view, the results show that ED-ConvNet-SA outperforms original SA in all three testing sets, which validates the robustness and effectiveness of ED-ConvNet.

Strategy 1 vs. 4: the Benefit of Image Partitioning. Different from traditional strategies querying at the image level, we introduce the image partitioning approach to generate dozens of patches that separates informative regions from the whole image. Strategy 4 achieves considerable improvement ($\sim$6% better in P-A & P-B and $\sim$4% better in Mixed) and even slightly better than ED-ConvNet-SA. This result answers our question in Sect. 1: the image-level query strategy is not good enough for the intensive cell image segmentation task. We can get better performance by selecting parts of valuable regions within an image rather than the whole.

Strategy 4 vs. 5: the Effect of Instance Diversity. We compare the proposed instance diversity criterion that reweights IU, to IU-only on image patches. Results show that our reweighting approach improves the Dice score to a certain extent. The reason is that the ID criterion avoids selecting too many uncertain image patches with high similarity, aka duplicate annotation. This helps refine the selection procedure to obtain more diverse patches.

Strategy 5 vs. 6 vs. 7: the Effect of Cluster-Wise Selection. In addition to image partitioning, we also explore the effect of clustering in the selection. Strategy 6 applies the mini-batch k-means algorithm and select image patches equally from clusters. It improves almost 1% Dice score compared to non-clustering strategy 5. As for the proposed cluster-instance double ranking algorithm in strategy 7, it has a further improvement for about 0.8%. This verifies the advantages of cluster importance estimation described in Eq. 8.

In general, FocAnnot is able to select less than half of the data to surpass FCN of full supervision. It also approximates (in Mixed) or even outperforms (in P-B)

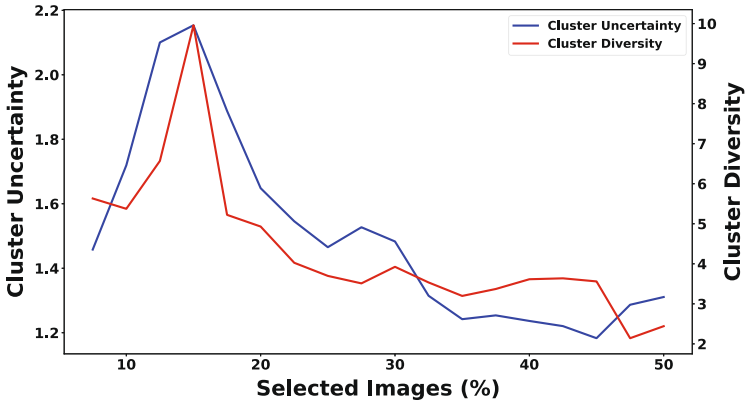


Fig. 5. Changes of averaged cluster uncertainty and cluster diversity during the selection.

ED-ConvNet using full training data. Specially, strategies 6 and 7 can achieve similar performance as ED-ConvNet-SA by selecting only 25% training data and obtain a good enough model with 30% data. All these experiments indicate that the combination of instance-level and cluster-level criteria is more potent than the image-level query strategy.

4.4 Visualization and Discussion

In Fig. 5, we visualize the changes in cluster uncertainty and cluster diversity during the selection on average. In the first few rounds, CU and CD are increasing and reach a peak when obtaining 15% of training data. This indicates that cluster-level criteria may not help much in the beginning. FocAnnot mainly depends on instance-level criteria to select valuable image patches at that time. After that, the segmentation model has learned enough knowledge to distinguish distinct patch descriptors. We can see that CU and CD decrease, which means cluster-level criteria begin to work at this stage. Representative image patches are continuously selected from groups. Hence clusters become denser and closer to each other. More concretely, CU and CD have a rapid drop between 15% to 25% and reach a relatively low position at 30%. This supports the previous conclusion, that is, the proposed query strategy can quickly improve segmentation accuracy and tends to be stable with around 30% training data.

FocAnnot has been proved effective and our query strategy is superior to the competitor. Moreover, we would like to investigate the model generalization in different parameter settings. As shown in Table 2, 12 parameter combinations are grouped into three categories based on patch sizes. The overlapped partition approach is always better than the seamless partition. We believe overlapped patches enrich local details of images and provide more valuable choices during the selection. Based on the overlapped setting, $k = 3$ is preferred except for two comparisons in P-A. Three clusters are enough to describe structures of cell

Table 2. Analysis of parameter combinations based on query strategy 7 with 50% selected training data. The best performances in each group are in bold. The underline indicates the setting outperforms all other combinations.

Parameters			Dice			VOE (%)		
l	ρ	k	Mixed	P-A	P-B	Mixed	P-A	P-B
64	0	3	0.857	0.805	0.854	24.40	31.51	24.72
64	0	5	0.852	0.806	0.843	24.92	31.47	26.51
64	0.5	3	**0.876**	0.836	**0.861**	**21.24**	27.36	**23.82**
64	0.5	5	0.866	**0.839**	0.855	22.76	**26.81**	24.55
128	0	3	0.856	0.802	0.848	24.17	32.00	25.63
128	0	5	0.867	0.815	0.869	22.69	30.00	22.60
128	0.5	3	**0.893**	0.848	**0.876**	**18.85**	25.40	**21.30**
128	0.5	5	0.887	**0.852**	0.872	19.80	**24.68**	22.02
256	0	3	0.875	0.792	0.855	21.49	32.55	24.66
256	0	5	0.870	0.820	0.860	22.22	29.24	23.67
256	0.5	3	**0.891**	**0.857**	**0.887**	**18.87**	**23.87**	**19.63**
256	0.5	5	0.886	0.843	0.864	19.81	26.30	23.31

objects in this dataset because their contours are less complicated. Too many categories would lead to ambiguous groups instead. The selection of patch size is critical. As we discussed in Sect. 3.2, a small patch size will lose content details and degrade model performance, for example, $l = 64$ vs. $l = 128$. On the other hand, a larger size brings limited benefits to the segmentation as well. Thus, a moderate size, i.e., $l = 128$, could be a good trade-off to reduce annotation cost with an acceptable accuracy at the same time.

In Fig. 6, we show the ten most valuable images selected by SA. Then, we mark ten patches per image in red, which are considered important based on FocAnnot. All these images are from the training set of Mixed, and the following names are their origin ID of GlaS. We observe that our query strategy prefers some interesting patterns:

- Most of the regions are located on the object boundaries, especially crossing two glands (#4, #50, #58, #84 in the training set). This pattern benefits the segmentation task on biomedical images with multiple and intensive objects. The model pays more attention to detailed differences among close glands.
- Contours with rough shape are identified and recommended for the annotation (#10, #19, #29 in the training set and #42 in test Part A). It is natural to select rare types of contours to improve model generalization and so our method does.
- Poorly differentiated epithelial cell nuclei and lumen in malignant glands are also useful to the model (#10, #29, #58 in the training set). This type of patch shows one of the most significant differences between benign and malignant glands in clinical practice.

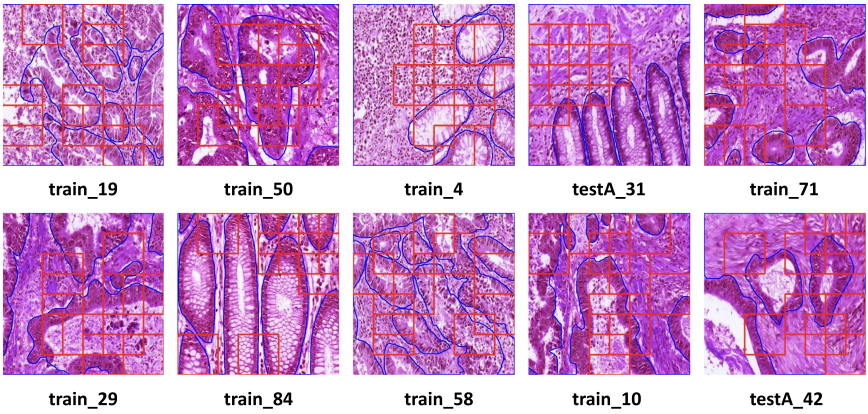


Fig. 6. Visualization of top 10 valuable images selected by SA. For each image, the first 10 queried patches by FocAnnot are marked as red rectangles. The blue curves show contours of glands.

According to these particular regions, more effective active learning methods can be developed by applying well-designed constraints.

5 Conclusion and Future Work

We have proposed a patch-wise active learning method, namely FocAnnot, for the intensive cell image segmentation problem based on an encoder-decoder convolutional network. The key idea is to partition images into patches and cluster them based on their high-level features. Hence, we can evaluate each patch separately to avoid duplicate annotation of similar cells within an image. Two criteria, i.e., cluster uncertainty and cluster diversity, are proposed to estimate the importance of each group. We also present an instance diversity criterion incorporated with the instance uncertainty to seek valuable data within clusters. The experimental results on a real-world cell-intensive dataset demonstrate that our method is able to reduce an additional 15% annotation cost or improve 2% segmentation performance compared to the competitor.

In the future, we would like to extend our query strategy in a multi-scale manner that combines local content with global knowledge. We are also interested in introducing object detection techniques to replace the hard partitioning approach with arbitrary size.

Acknowledgment. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFC1001703, in part by the National Natural Science Foundation of China under Grant 61825205, Grant 61772459, and in part by the National Science and Technology Major Project of China under Grant 50-D36B02-9002-16/19.

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2481–2495 (2017)
2. Bonnin, A., Borràs, R., Vitrià, J.: A cluster-based strategy for active learning of RGB-D object detectors. In: ICCV Workshops, pp. 1215–1220 (2011)
3. Chen, B.k., Gong, C., Yang, J.: Importance-aware semantic segmentation for autonomous driving system. In: IJCAI, pp. 1504–1510 (2017)
4. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2018)
5. Chen, L.C., Yang, Y., Wang, J., Xu, W., Yuille, A.L.: Attention to scale: Scale-aware semantic image segmentation. In: CVPR, pp. 3640–3649 (2016)
6. Chyzhyk, D., Dacosta-Aguayo, R., Mataró, M., Graña, M.: An active learning approach for stroke lesion segmentation on multimodal MRI data. Neurocomputing **150**, 26–36 (2015)
7. Dou, Q., Chen, H., Jin, Y., Yu, L., Qin, J., Heng, P.-A.: 3D deeply supervised network for automatic liver segmentation from CT volumes. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 149–157. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_18
8. Dutt Jain, S., Grauman, K.: Active image segmentation propagation. In: CVPR, pp. 2864–2873 (2016)
9. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: ICCV, pp. 2650–2658 (2015)
10. Gal, Y., Islam, R., Ghahramani, Z.: Deep Bayesian active learning with image data. In: ICML, pp. 1183–1192 (2017)
11. Hoogi, A., Subramaniam, A., Veerapaneni, R., Rubin, D.L.: Adaptive estimation of active contour parameters using convolutional neural networks and texture analysis. IEEE Trans. Med. Imaging **36**(3), 781–791 (2017)
12. Iglesias, J.E., Konukoglu, E., Montillo, A., Tu, Z., Criminisi, A.: Combining generative and discriminative models for semantic segmentation of CT scans via active learning. In: Székely, G., Hahn, H.K. (eds.) IPMI 2011. LNCS, vol. 6801, pp. 25–36. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-22092-0_3
13. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning, pp. 448–456 (2015)
14. Konyushkova, K., Sznitman, R., Fua, P.: Introducing geometry in active learning for image segmentation. In: ICCV, pp. 2974–2982 (2015)
15. Konyushkova, K., Sznitman, R., Fua, P.: Learning active learning from data. In: NeurIPS, pp. 4228–4238 (2017)
16. Lin, C.H., Mausam, M., Weld, D.S.: Re-active learning: active learning with relabeling. In: AAAI (2016)
17. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR, pp. 3431–3440 (2015)

18. Mahapatra, D., Bozorgtabar, B., Thiran, J.-P., Reyes, M.: Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11071, pp. 580–588. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_65
19. Mahapatra, D., Buhmann, J.M.: Visual saliency based active learning for prostate MRI segmentation. In: Zhou, L., Wang, L., Wang, Q., Shi, Y. (eds.) MLMI 2015. LNCS, vol. 9352, pp. 9–16. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24888-2_2
20. Mahapatra, D., et al.: Active learning based segmentation of Crohn's disease using principles of visual saliency. In: ISBI, pp. 226–229 (2014)
21. Mansoor, A., et al.: Deep learning guided partitioned shape model for anterior visual pathway segmentation. IEEE Trans. Med. Imaging **35**(8), 1856–1865 (2016)
22. Möller, T., Nillsen, I., Nattkemper, T.W.: Active learning for the classification of species in underwater images from a fixed observatory. In: ICCV (2017)
23. Nguyen, H.T., Smeulders, A.: Active learning using pre-clustering. In: ICML, p. 79 (2004)
24. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
25. Sirinukunwattana, K., et al.: Gland segmentation in colon histology images: the glas challenge contest. Med. Image Anal. **35**, 489–502 (2017)
26. Song, Y., et al.: Accurate cervical cell segmentation from overlapping clumps in pap smear images. IEEE Trans. Med. Imaging **36**(1), 288–300 (2017)
27. Volpi, M., Tuia, D.: Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. IEEE Trans. Geosci. Remote Sens. **55**(2), 881–893 (2017)
28. Xing, F., Xie, Y., Yang, L.: An automatic learning-based framework for robust nucleus segmentation. IEEE Trans. Med. Imaging **35**(2), 550–566 (2016)
29. Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D.Z.: Suggestive annotation: a deep active learning framework for biomedical image segmentation. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 399–407. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_46
30. Yang, Y., Ma, Z., Nie, F., Chang, X., Hauptmann, A.G.: Multi-class active learning by uncertainty sampling with diversity maximization. Int. J. Comput. Vis. **113**(2), 113–127 (2015)
31. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
32. Yu, L., Chen, H., Dou, Q., Qin, J., Heng, P.A.: Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE Trans. Med. Imaging **36**(4), 994–1004 (2017)
33. Yu, L., Yang, X., Chen, H., Qin, J., Heng, P.A.: Volumetric ConvNets with mixed residual connections for automated prostate segmentation from 3D MR images. In: AAAI, pp. 66–72 (2017)
34. Zhou, Z., Shin, J., Zhang, L., Gurudu, S., Gotway, M., Liang, J.: Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally. In: CVPR, pp. 7340–7349 (2017)