Learning to segment microscopy images with lazy labels

Anonymous ECCV submission

Paper ID 100

Abstract. The need for labour intensive pixel-wise annotation is a major limitation of many fully supervised learning methods for segmenting bioimages that can contain numerous object instances with thin separations. In this paper, we introduce a deep convolutional neural network for microscopy image segmentation. Annotation issues are circumvented by letting the network being trainable on coarse labels combined with only a very small number of images with pixel-wise annotations. We call this new labelling strategy 'lazy' labels. Image segmentation is stratified into three connected tasks: rough inner region detection, object separation and pixel-wise segmentation. These tasks are learned in an end-to-end multi-task learning framework. The method is demonstrated on two microscopy datasets, where we show that the model gives accurate segmentation results even if exact boundary labels are missing for a majority of annotated data. It brings more flexibility and efficiency for training deep neural networks that are data hungry and is applicable to biomedical images with poor contrast at the object boundaries or with diverse textures and repeated patterns.

Keywords: Microscopy images, Multi-task learning, Convolutional neural networks, Image segmentation

1 Introduction

Image segmentation is a crucial step in many microscopy image analysis problems. It has been an active research field in the past decades. Deep learning approaches play an increasingly important role and have become state-of-the-art in various segmentation tasks [12, 17, 21, 27, 40]. However, the segmentation of microscopy images is very challenging not only due to the fact that these images are often of low contrast with complex instance structures, but also because of the difficulty in obtaining ground truth pixel-wise annotations [2, 15] which hinders the applications of recent powerful but data-hungry deep learning techniques.

In this paper, we propose a simple yet effective multi-task learning approach for microscopy image segmentation. We address the problem of finding segmentation with accurate object boundaries from mainly rough labels. The labels are all pixel-wise and contain considerable information about individual objects, but they are created relatively easily. The method is different from pseudo labelling (PL) approaches, which generate fake training segmentation masks from coarse labels and may induce a bias in the masks for microscopy data.

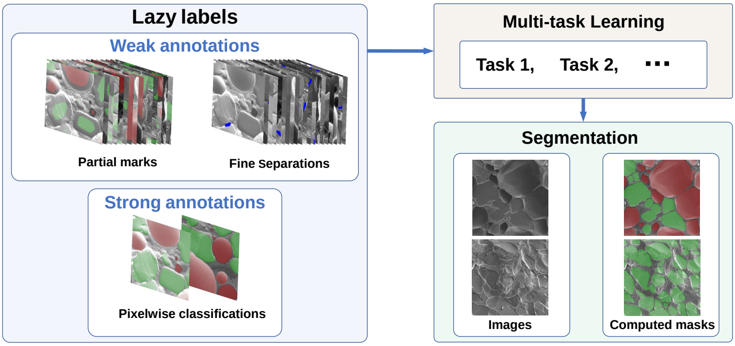


Fig. 1. Multi-task learning for image segmentation with lazy labels. The figure uses Scanning Electron Microscopy (SEM) images of food microstructures as an example and demonstrates a segmentation problem of three classes, namely air bubbles (green), ice crystals (red) and background respectively. Most of the training data are weak annotations containing (i) partial marks of ice crystals and/or air bubbles instances and (ii) fine separation marks of boundaries shared by different instances. Only a *few* strongly annotated images are used. On the bottom right SEM images and their corresponding segmentation outputs from the learned multi-task model are shown.

To circumvent the need for a massive set of ground truth segmentation masks, we rather develop a segmentation approach that we split into three relevant tasks: detection, separation and segmentation (cf. Figure 1). Doing so, we obtain a weakly supervised learning approach that is trained with what we call "lazy" labels. These lazy labels contain a lot of coarse annotations of class instances, together with a few accurately annotated images that can be obtained from the coarse labels in a semi-automated way. Contrary to PL approaches, only a very limited number of accurate annotation are considered. In the following, we will refer to weak (resp. strong) annotations for coarse (resp. accurate) labels and denote them as WL (resp. SL).

We reformulate the segmentation problem into several more tractable tasks that are trainable on less expensive annotations, and therefore reduce the overall annotation cost. The *first task* detects and classifies each object by roughly determining its inner region with an under-segmentation mask. Instance counting can be obtained as a by-product of this task. As the main objective is instance detection, exact labels for the whole object or its boundary are not necessary at this stage. We use instead weakly annotated images in which a rough region inside each object is marked, cf. the most top left part of Figure 1. For segmentation problems with a dense population of instances, such as the food components (see e.g., Figure 1), cells [13, 33], glandular tissue, or people in a crowd [42], separating objects sharing a common boundary is a well known challenge. We can optionally perform a *second task* that focuses on the separation of instances that are connected without a clear boundary dividing them. Also for this task we rely on WL to reduce the burden of manual annotations: touching interfaces are specified with rough scribbles, cf. top left part of Figure 1. Note that this task is suitable for applications with instances that are occasionally connected without clear boundaries. One can alternatively choose to have fewer labelled samples in this task if the annotation cost per sample is higher. The *third task* finally tackles pixel-wise classification of instances. It requires strong annotations that are accurate up to the object boundaries. Thanks to the information brought by weak annotations, we here just need a very small set of accurate segmentation masks, cf. bottom left part of Figure 1. To that end, we propose to refine some of the coarse labels resulting from task 1 using a semi-automatic segmentation method which requires additional manual intervention.

The three tasks are handled by a single deep neural network and are jointly optimized using a cross entropy loss. In this work we use a network architecture inspired by U-net [33] which is widely used for segmenting objects in microscopy images. While all three tasks share the same contracting path, we introduce a new multi-task block for the expansive path. The network has three outputs and is fed with a combination of WL and SL described above. Obtaining accurate segmentation labels for training is usually a hard and time consuming task. We here demonstrate that having exact labels for a small subset of the whole training set does not degrade training performances. We evaluate the proposed approach on two microscopy image datasets, namely the segmentation of SEM images of food microstructure and stained histology images of glandular tissues.

The contributions of the paper are threefold. (1) We propose a decomposition of the segmentation problems into three tasks and a corresponding user friendly labelling strategy. (2) We develop a simple and effective multi-task learning framework that learns directly from the coarse and strong manual labels and is trained end-to-end. (3) Our approach outperforms the pseudo label approaches on the microscopy image segmentation problems being considered.

2 Related Work

In image segmentation problems, one needs to classify an image at pixel level. It is a vast topic with a diversity of algorithms and applications being considered, including traditional unsupervised methods like k-means clustering [29] that splits the image into homogeneous regions according to image low level features, curve evolution based methods like snakes [7], graph-cut based methods [5,24,34], just to name a few. Interactive approaches like snakes or Grabcut enable getting involved users' knowledge by means of initializing regions or putting constraints on the segmentation results. For biological imaging, the applications of biological prior knowledge, such as shape statistics [14], semantic information [24] and atlas [10], is effective for automatic segmentation approaches.

2.1 Deep neural networks for segmentation

In the last years, numerous deep convolutional neural network (DCNN) approaches have been developed for segmenting complex images, especially in the semantic setting. In this work, we rely more specifically on fully convolutional networks (FCN) [28], that replace the last few fully connected layers of a conventional classification network by up-sampling layers and convolutional layers, to preserve spatial information. FCNs have many variants for semantic segmentation. The DeepLab [9] uses a technique called atrous convolution to handle spatial information together with a fully connected conditional random field (CRF) [8] for refining the segmentation results. Fully connected CRF can be used as post-processing or can be integrated into the network architecture, allowing for end-to-end training [46].

One type of FCNs commonly used in microscopy and biomedical image segmentation are encoder-decoder networks [1,33]. They have multiple up-sampling layers for better localizing boundary details. One of the most well-known models is the U-net [33]. It is a fully convolutional network made of a contracting path, which brings the input images into very low resolution features with a sequence of down-sampling layers, and an expansive path that has an equal amount of up-sampling layers. At each resolution scale, the features on the contracting path are merged with the corresponding up-sampled layers via long skip connections to recover detailed structural information, e.g., boundaries of cells, after down-sampling.

2.2 Weakly supervised learning and multi-task learning

Standard supervision for semantic segmentation relies on a set of image and ground truth segmentation pairs. The learning process contains an optimization step that minimizes the distance between the outputs and the ground truths. There has been a growing interest in weakly supervised learning, motivated by the heavy cost of pixel-level annotation needed for fully supervised methods. Weakly supervised learning uses weak annotations such as image-level labels [17,25,30–32,36,47], bounding boxes [21,35], scribbles [26] and points [3].

Many weakly supervised deep learning methods for segmentation are built on top of a classification network. The training of such networks may be realized using segmentation masks explicitly generated from weak annotations [21,25,40,43,44]. The segmentation masks can be improved recursively, which involves several rounds of training of the segmentation network [11,19,43]. Composite losses from some predefined guiding principles are also proposed as supervision from the weak signals [23,23,38].

Multi-task techniques aim to boost the segmentation performance via learning jointly from several relevant tasks. Tailored to the problems and the individual tasks of interest, deep convolutional networks have been designed, for example, the stacked U-net for extracting roads from satellite imagery [39], the two stage 3D U-net framework for 3D CT and MR data segmentation [41], encoder-decoder networks for depth regression, semantic and instance segmentation [20], or the cascade multi-task network for the segmentation of building footprint [4].

Segmentation of microscopy and biomedical images. Various multi-task deep learning methods have been developed for processing microscopy images and biomedical images. An image level lesion detection task [32] is investigated for the segmentation of retinal red/bright lesions. The work [30] considers to jointly segment and classify brain tumours. A deep learning model is developed in [48] to simultaneously predict the segmentation maps and contour maps for pelvic

CT images. In [15], an auxiliary task that predicts centre point vectors for nuclei segmentation in 3D microscopy images is proposed. Denoising tasks, which aims to improve the image quality, can also be integrated for better microscopy image segmentation [6].

In this work, the learning is carried out in a weakly supervised fashion with weak labels from closely related tasks. Nevertheless, the proposed method exploits cheap and coarse pixel-wise labels instead of very sparse image level annotations and is more specialized in distinguishing the different object instances and clarifying their boundaries in microscopy images. The proposed method is completely data-driven and it significantly reduces the annotation cost needed by standard supervision. We aim at obtaining segmentation with accurate object boundaries from mainly coarse pixel-wise labels.

3 Multi-task learning framework

The objective of fully supervised learning for segmentation is to approximate the conditional probability distribution of the segmentation mask given the image. Let $\boldsymbol{s}^{(3)}$ be the ground truth segmentation mask and I be the image, then the segmentation task aims to estimate $p(\boldsymbol{s}^{(3)} \mid I)$ based on a set of sample images $\mathcal{I} = \{I_1, I_2, \cdots, I_n\}$ and the corresponding labels $\{\boldsymbol{s}_1^{(3)}, \boldsymbol{s}_2^{(3)}, \cdots, \boldsymbol{s}_n^{(3)}\}$. The set $\mathcal{I}$ is randomly drawn from an unknown distribution. In our setting, having the whole set of segmentation labels $\{\boldsymbol{s}_i^{(3)}\}_{1,\cdots,n}$ is impractical, and we introduce two auxiliary tasks for which the labels can be more easily generated to achieve an overall small cost on labelling.

For a given image $I \in \mathcal{I}$, we denote as $\boldsymbol{s}^{(1)}$ the rough instance detection mask, and $\boldsymbol{s}^{(2)}$ a map containing some interfaces shared by touching objects. All labels $\boldsymbol{s}^{(1)}, \boldsymbol{s}^{(2)}, \boldsymbol{s}^{(3)}$ are represented in one-hot vectors. For the first task, the contours of the objects are not treated carefully, resulting in a coarse label mask $\boldsymbol{s}^{(1)}$ that misses most of the boundary pixels, cf left of Figure 1. In the second task, the separation mask $\boldsymbol{s}^{(2)}$ only specifies connected objects without clear boundaries rather than their whole contours. Let $\mathcal{I}_k \subset \mathcal{I}$ denote the subset of images labelled for task k $(k = 1, 2, 3)$. As we collect a different amount of annotations for each task, the number of annotated images $|\mathcal{I}_k|$ may not be the same for different k. Typically the number of images with strong annotations satisfies $|\mathcal{I}_3| \ll n$, as the annotation cost per sample is higher.

The set of samples in $\mathcal{I}_3$ for segmentation being small, the computation of an accurate approximation of the true probability distribution $p(\boldsymbol{s}^{(3)} \,|\, I)$ is a challenging issue. Given that much more samples of $\boldsymbol{s}^{(1)}$ and $\boldsymbol{s}^{(2)}$ are observed, it is simpler to learn the statistics of these *weak* labels. Therefore, in a multi-task learning setting, one also aims at approximating the conditional probabilities $p(\boldsymbol{s}^{(1)} \,|\, I)$ and $p(\boldsymbol{s}^{(2)} \,|\, I)$ for the other two tasks, or the joint probability $p(\boldsymbol{s}^{(1)}, \boldsymbol{s}^{(2)}, \boldsymbol{s}^{(3)} \,|\, I)$. The three tasks can be related to each other as follows. First, by the definition of the detection task, one can see that $p(\boldsymbol{s}^{(3)} = z \,|\, \boldsymbol{s}^{(1)} = x) = 0$ for x and z satisfying $x_{i,c} = 1$ and $z_{i,c} = 0$ for some pixel i and class c other than the background. Next, the map of interfaces $\boldsymbol{s}^{(2)}$ indicates small gaps between two connected instances, and is therefore a subset of boundary pixels of the mask $\boldsymbol{s}^{(3)}$.

Let us now consider the probabilities given by the models $p(\boldsymbol{s}^{(k)} \mid I; \boldsymbol{\theta})$ $(k = 1, 2, 3)$ parameterized by $\boldsymbol{\theta}$, that will consist of network parameters in our setting. We do not optimize $\boldsymbol{\theta}$ for individual tasks, but instead consider a joint probability $p(\boldsymbol{s}^{(1)}, \boldsymbol{s}^{(2)}, \boldsymbol{s}^{(3)} \mid I; \boldsymbol{\theta})$, so that the parameter $\boldsymbol{\theta}$ is shared among all tasks. Assuming that $\boldsymbol{s}^{(1)}$ (rough under-segmented instance detection) and $\boldsymbol{s}^{(2)}$ (a subset of shared boundaries) are conditionally independent given image I, and if the samples are i.i.d., we define the maximum likelihood (ML) estimator for $\boldsymbol{\theta}$ as

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \sum_{I \in \mathcal{I}} \left(\log p\left(\boldsymbol{s}^{(3)} \mid \boldsymbol{s}^{(1)}, \boldsymbol{s}^{(2)}, I; \boldsymbol{\theta}\right) + \sum_{k=1}^{2} \log p\left(\boldsymbol{s}^{(k)} \mid I; \boldsymbol{\theta}\right)\right). \tag{1}$$

The set $\mathcal{I}_3$ may not be evenly distributed across $\mathcal{I}$, but we assume that it is generated by a fixed distribution as well. Provided that the term $\{p(\boldsymbol{s}^{(3)} \mid \boldsymbol{s}^{(1)}, \boldsymbol{s}^{(2)}, I)\}_{I \in \mathcal{I}}$ can be approximated correctly by $p(\boldsymbol{s}^{(3)} \,|\, \boldsymbol{s}^{(1)}, \boldsymbol{s}^{(2)}, I; \boldsymbol{\theta})$ even if $\boldsymbol{\theta}$ is computed without $\boldsymbol{s}^{(3)}$ specified for $\mathcal{I} \backslash \mathcal{I}_3$, then

$$\sum_{I \in \mathcal{I}} \log p\left(\boldsymbol{s}^{(3)} \,|\, \boldsymbol{s}^{(1)}, \boldsymbol{s}^{(2)}, I; \boldsymbol{\theta}\right) \propto \sum_{I \in \mathcal{I}_3} \log p\left(\boldsymbol{s}^{(3)} \,|\, \boldsymbol{s}^{(1)}, \boldsymbol{s}^{(2)}, I; \boldsymbol{\theta}\right). \tag{2}$$

Finally assuming that the segmentation mask does not depend on $\boldsymbol{s}^{(1)}$ or $\boldsymbol{s}^{(2)}$ given $I \in \mathcal{I}_3$, and if $|\mathcal{I}_1|, |\mathcal{I}_2|$ are large enough, then from Equations (1), and (2), we approximate the ML estimator by

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_{k=1}^{3} \sum_{I \in \mathcal{I}_k} \alpha_k \log p\left(\boldsymbol{s}^{(k)} \mid I; \boldsymbol{\theta}\right) \tag{3}$$

in which $\alpha_1, \alpha_2, \alpha_3$ are non negative constants.

3.1 Loss function

Let the outputs of the approximation models be denoted respectively by $h_{\boldsymbol{\theta}}^{(1)}(I)$, $h_{\boldsymbol{\theta}}^{(2)}(I)$, and $h_{\boldsymbol{\theta}}^{(3)}(I)$, with $\left[h_{\boldsymbol{\theta}}^{(k)}(I)\right]_{i,c}$ the estimated probability of pixel i to be in class c of task k. For each task k, the log likelihood function related to the label $\boldsymbol{s}^{(k)}$ writes

$$\log p\left(\boldsymbol{s}^{(k)} \mid I; \boldsymbol{\theta}\right) = \sum_{i} \sum_{c \in C_k} s_{i,c}^{(k)} \log \left[h_{\boldsymbol{\theta}}^{(k)}(I)\right]_{i,c}, \quad k = 1, 2, 3, \tag{4}$$

in which $s_{i,c}^{(k)}$ denotes the element of the label $\boldsymbol{s}^{(k)}$ at pixel i for class c and C_k is the set of classes for task k. For example, for SEM images of ice cream (see details in section 4.1), we have three classes including air bubbles, ice crystals and the rest (background or parts of the objects ignored by the weak labels), so $C_1, C_3 = \{1, 2, 3\}$. For the separation task, there are only two classes for pixels (belonging or not to a touching interface) and $C_2 = \{1, 2\}$. According to Equation (3), the network is trained by minimizing the weighted cross entropy loss:

$$L(\boldsymbol{\theta}) = -\sum_{I \in \mathcal{I}} \sum_{k=1}^{3} \alpha_k \mathbb{1}_{\mathcal{I}_k}(I) \log p\left(\boldsymbol{s}^{(k)} \mid I; \boldsymbol{\theta}\right), \tag{5}$$

Here $\mathbb{1}_{\mathcal{I}_k}(\cdot)$ is an indicator function which is 1 if $I \in \mathcal{I}_k$ and 0 otherwise.

3.2 Multi-task Network

We follow a convolutional encoder-decoder network structure for multi-task learning. The network architecture is illustrated in Figure 2. As an extension of the U-net structure for multiple tasks, we only have one contracting path that encodes shared features representation for all the tasks. On the expansive branch, we introduce a multi-task block at each resolution to support different learning purposes (blue blocks in Figure 2). Every multi-task block runs three paths, with three inputs and three corresponding outputs, and it consists of several sub-blocks.

In each multi-task block, the detection task (task 1) and the segmentation task (task 3) have a common path similar to the decoder part of the standard U-net. They share the same weights and use the same concatenation with feature maps from contracting path via the skip connections. However, we insert an additional residual sub-block for the segmentation task. The residual sub-block provides extra network parameters to learn information not known from the detection task, *e.g.* object boundary localization. The path for the separation task (task 2) is built on the top of detection/segmentation ones. It is also a U-net decoder block structure, but the long skip connections start from the sub-blocks of the detection/segmentation paths instead of the contracting path. The connections extract higher resolution features from the segmentation task and use them in the separation task.

To formulate the multi-task blocks, let $\boldsymbol{x}_l$ and $\boldsymbol{z}_l$ denote respectively the output of the detection path and segmentation path at the multi-task block l, and let $\boldsymbol{c}_l$ be the feature maps received from the contracting path with the skip connections. Then for task 1 and task 3 we have

$$\begin{cases} \boldsymbol{x}_{l+1} = F_{W_l}(\boldsymbol{x}_l, \boldsymbol{c}_l), \\ \boldsymbol{z}_{l+\frac{1}{2}} = F_{W_l}(\boldsymbol{z}_l, \boldsymbol{c}_l), \quad \boldsymbol{z}_{l+1} = \boldsymbol{z}_{l+\frac{1}{2}} + F_{W_{l+\frac{1}{2}}}(\boldsymbol{z}_{l+\frac{1}{2}}), \end{cases} \tag{6}$$

in which $W_l, W_{l+1/2} \in \boldsymbol{\theta}$ are subsets of network parameters and F_{W_l}, $F_{W_{l+\frac{1}{2}}}$ are respectively determined by a sequence of layers of the network (cf. small grey blocks on the right of Figure 2). For task 2 the output at l^{th} block $\boldsymbol{y}_{l+1}$ is computed as $\boldsymbol{y}_{l+1} = G_{\tilde{W}_l}(\boldsymbol{z}_{l+1}, \boldsymbol{y}_l)$ with additional network parameters $\tilde{W}_l \in \boldsymbol{\theta}$. Finally, after the last multi-task block, softmax layers are added, outputting a probability map for each task.

Implementation details. We implement a multi-task U-net with 6 levels of spatial resolution and input images of size 256×256. A sequence of down-sampling via max-pooling with pooling size 2×2 is used for the contracting path of the network. Different from the conventional U-net [33], each small grey block (see Figure 2) consists of a convolution layer and a batch normalization [18], followed by a leaky ReLU activation with a leakiness parameter 0.01. The same setting is also applied to grey sub-blocks of the 4 multi-task blocks. On the expansive path of the network, feature maps are up-sampled (with factor 2×2) by bilinear interpolation from a low resolution multi-task block to the next one.

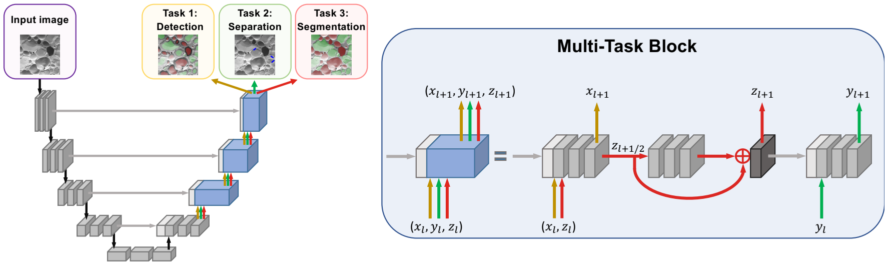


Fig. 2. Architecture of the multi-task U-net. The left part of the network is a contracting path similar to the standard U-net. For multi-task learning, we construct several expansive paths with specific multi-task blocks. At each resolution, task 1 (Detection in yellow) and task 3 (Segmentation in red) run through a common sub-block, but the red path learns an additional residual to better localize object boundaries. Long skip connections with the layers from contracting path are built for yellow/red paths via concatenation. Task 2 (Separation, in green) mainly follows a separated expansive path, with its own up-sampled blocks. A link with the last layer of task 3 is added via a skip connection in order to integrate accurate boundaries in the separation task.

3.3 Methods for lazy labels generation

We now explain our strategy for generating all the lazy annotations that are used for training. We introduce our method with a data set of ice cream SEM images but any other similar microscopy datasets could be used. Typical images of ice cream samples are shown in the top row of the left part of Figure 3. The segmentation problem is challenging since the images contain densely distributed small object instances (*i.e.*, air bubble and ice crystals), and poor contrast between the foreground and the background. The sizes of the objects can vary significantly in a single sample. Textures on the surfaces of objects also appear.

As a first step, scribble-based labelling is applied to obtain detection regions of air bubbles and ice crystals for task 1. This can be done in a very fast way as no effort is put on the exact object boundaries. We adopt a lazy strategy by picking out an inner region for each object in the images (see *e.g.*, the second row of the left part of Figure 3). Though one could get these rough regions as accurate as possible, we delay such refinement to task 3, for better efficiency of the global annotation process. Compared to the commonly used bounding box annotations in computer vision tasks, these labels give more confidence for a particular part of the region of interest.

In the second step, we focus on tailored labels for those instances that are close one to each other (task 2), without a clear boundary separating them. Again, we use scribbles to mark their interface. Examples for such annotations are given in Figure 3 (top line, right part) The work can be carried out efficiently especially when the target scribbles have a sparse distribution. On the other hand, as no labelling is needed for the objects that are well separated, we can collect sufficient labelled images in a limited amount of time and cover the complex ice cream sample conditions. Lazy manual labelling of tasks 1 and 2 are done independently.

It follows the assumption made in Section 3 that $\boldsymbol{s}^{(1)}$ and $\boldsymbol{s}^{(2)}$ are conditionally independent given image I.

The precise labels for task 3 are created using interactive segmentation tools. Starting from the rough (inner) regions of task 1, a natural idea is to let these regions grow and stop when the boundaries are reached. This can be done with geodesic active contours [7]. Unfortunately, such a method fails to capture sharp corners and the contour evolution tends to ignore boundaries with low contrast. The annotation then requires frequent and time consuming user interaction. Instead, we use Grabcut [16, 34] a graph-cut based method. The initial labels obtained from the first step give a good guess of the whole object regions. The Grabcut works well on isolated objects. However, it gives poor results when the objects are close to each other and have boundaries with inhomogeneous colors. As corrections may be needed for each image, only a few images of the whole dataset are processed. A fully segmented example is shown in the last row of Figure 3.

4 Experiments

In this section, we demonstrate the performance of our approach using two microscopy image datasets. For both, we use strong , (SL) and weak labels (WL). We prepare the labels and design the network as described in Section 3.

4.1 Segmenting SEM images of ice cream

Scanning Electron Microscopy (SEM) constitutes the state-of-the-art for analysing food microstructures as it enables the efficient acquisition of high quality images for food materials, resulting into a huge amount of image data available for analysis. However, to better delineate the microstructures and provide exact statistical information, the partition of the images into different structural components and instances is needed. The structures of food, especially soft solid materials, are usually complex which makes automated segmentation a difficult task. Some SEM images of ice cream in our dataset are shown on the bottom right of Figure 1. A typical ice cream sample consists of air bubbles, ice crystals and a concentrated unfrozen solution. In most situations, the air bubbles and ice crystals appear as foam in the images, while the solution fills the gaps between them. We treat the solution as the background and aim at detecting and computing a pixel-wise classification for each air bubbles and ice crystals instances.

The set of ice-cream SEM dataset consists of 38 wide field-of-view and high resolution images that are split into three sets (53% for training, 16% for validation and 31% testing respectively). Each image contains a rich set of instances with an overall number of instances around 13300 for 2 classes (ice crystals and air bubbles). For comparison, the PASCAL VOC 2012 dataset has 27450 objects in total for 20 classes.

For training the network, data augmentation is applied to prevent over-fitting. The size of the raw images is 960×1280. They are rescaled and rotated randomly, and then cropped into an input size of 256×256 for feeding the network. Random

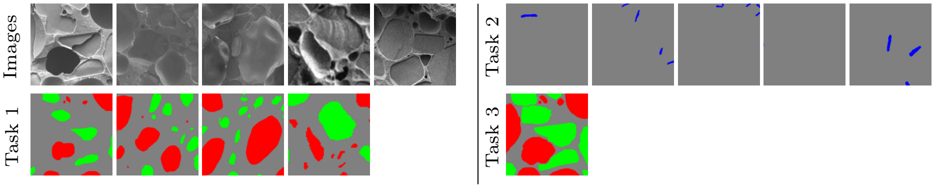


Fig. 3. Example of annotated images. Some of the annotations are not shown because the images are not labelled for the associated tasks. The red color and green color are for air bubbles and ice crystals, respectively. The blue curves in Task 2 are labels for interfaces of touching objects.

flipping is also performed during training. The network is trained using Adam optimizer [22] with a learning rate $r = 2 \times 10^{-4}$ and a batch size of 16.

In the inference phase, the network outputs for each patch a probability map of size 256×256. The patches are then aggregated to obtain a probability map for the whole image. In general, the pixels near the boundaries of each patch are harder to classify. We thus weight the spatial influence of the patches with a Gaussian kernel to emphasize the network prediction at patch center.

Table 1. Dice scores of segmentation results on the test images of SEM images of ice cream dataset.

The models	air bubbles	ice crystals	Overall
U-net on WL	0.725	0.706	0.716
U-net on SL	0.837	0.794	0.818
PL approach	0.938	0.909	0.924
Multi-task U-net	**0.953**	**0.931**	**0.944**

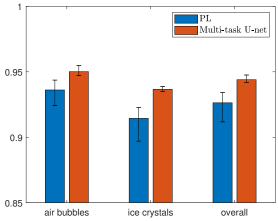
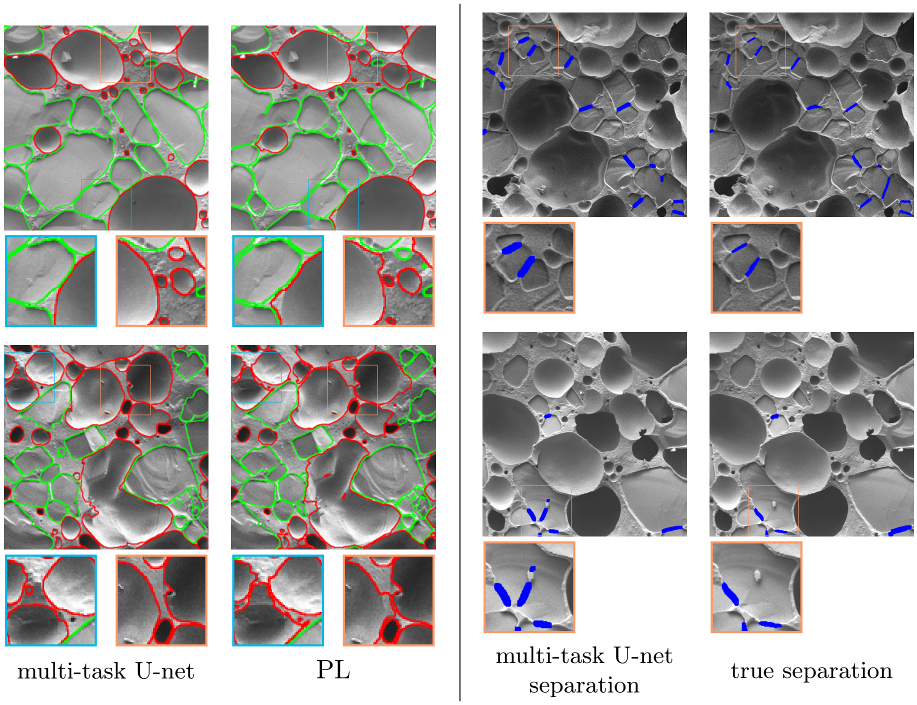


Fig. 4. The error bars for the PL and multi-task U-net. The top of each box represent the mean of the scores over 8 different experiments, the minimum and maximum of which are indicated by the whiskers

We now evaluate the multi-task U-net and compare it to the traditional single task U-net. The performance of each model is tested on 12 wide FoV images, and average results are shown in Table 1. In the table, the dice score for a class c is defined as $d_c = 2\sum_i x_{i,c}y_{i,c}/(\sum_i x_{i,c} + \sum_i y_{i,c})$where x is the computed segmentation mask and y the ground truth.

We train a single task U-net (*i.e.*, without the multi-task block) on the weakly labelled set (task 1), with the 15 annotated images. The single task U-net on weak annotations gives an overall dice score at 0.72, the lowest one among the three other methods tested. One reason for the low accuracy of the single task U-net on weak (inaccurate) annotations is that in the training labels, the object boundaries are mostly ignored. Hence the U-net is not trained to recover them, leaving large parts of the object not recognized. Second, we consider strong annotations as training data, without the data of the other tasks, *i.e.* only 2 images with accurate segmentation masks are used. The score of the U-net trained on SL is only 0.82, which is significantly lower than the 0.94 obtained by our multi-task network.



multi-task U-net

PL

multi-task U-net separation

true separation

Fig. 5. Segmentation and separation results (best view in color). First two columns: the computed contours are shown in red for air bubbles and green for ice crystals. While multi-task U-net and PL supervised network both have good performance, PL misclassifies the background near object boundaries. Last two columns: Examples of separation by the multi-task U-net and the ground truth.

We also compare our multi-task U-net results with one of the major weakly supervised approaches that make use of pseudo labels (PL) (see e.g., [19, 21]). In these approaches, the pseudo segmentation masks are created from WLs and are used to feed a segmentation network. Following the work of [21], we use the Grabcut method to create the PLs from the partial masks of task 1. For the small subset of images that are strongly annotated, the full segmentation masks are used instead of PLs. The PLs are created without human correction, and then used for feeding the segmentation network. Here we use the single task U-net for baseline comparisons.

Our multi-task network outperforms the PL approach as shown in Table 1. Figure 4 displays the error bars for the two methods with dice scores collected from 8 different runs. The performance of the PL method relies on the tools used for pseudo segmentation mask generations. If the tools create bias in the pseudo labels, then the learning will be biased as well, which is the case in this example. The images in the left part of Figure 5 show that the predicted label of an object tends to merge with some background pixels when there are edges of another object nearby.

Besides the number of pixels that are correctly classified, the separation of touching instances is also of interest. In addition to the dice scores in Table 1, we study the learning performance of our multi-task network on task 2, which specializes in the separation aspect. The test results on the 12 images give an overall precision of 0.70 of the detected interfaces, while 0.82 of the touching objects are recognized. We show some examples of computed separations and ground truth in the right part of Figure 5.

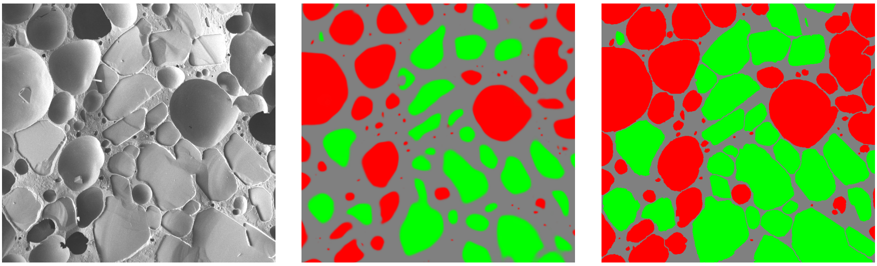


Fig. 6. The image (left), the inaccurate label predicted by the network for the detection task (middle), and the ground truth segmentation mask (right). The red and green colors on the middle and right images stand for air bubbles and ice crystals respectively.

For the detection task, the network predicts a probability map for the inner regions of the object instances. An output of the network is shown in Figure 6. With partial masks as coarse labels for this task, the network learns to identify the object instances.

Table 2. Comparison between the two methods under similar annotation time budgets. In each budget, two different combinations of SL and WL that take similar annotation time are used. Dice score is reported for each budget.

	methods	labels	dice score
Annotation budget 1	multi-task	10% SL + WL	**0.944**
	single task	20% SL	0.882
Annotation budget 2	multi-task	20% SL + WL	**0.948**
	single task	30% SL	0.913
Annotation budget 3	multi-task	50% SL + WL	**0.949**
	single task	75% SL	0.940

We finally consider the work of [3] that investigates the cost related to different types of annotations. Based on the data reported [3] and our estimated annotation time, the collecting of WL for detection is considerably (more than 6x) faster than obtaining strong labels SL. For a fair comparison with the baseline U-net we use a larger ratio of SL for the single task learning accordingly (since no WL is used here), and the results are reported in Table 2. The WL in this table contains 75% labels for the detection task and 100% labels for the separation task. From budget 1 to budget 3, we increase the amount of labels (that means more annotation time is needed) in the training data. The proposed method outperforms the U-net by a large margin on similar annotation time budgets, and we observe that additional SL after the first 10% do not help significantly.

4.2 Gland segmentation on H&E-stained images

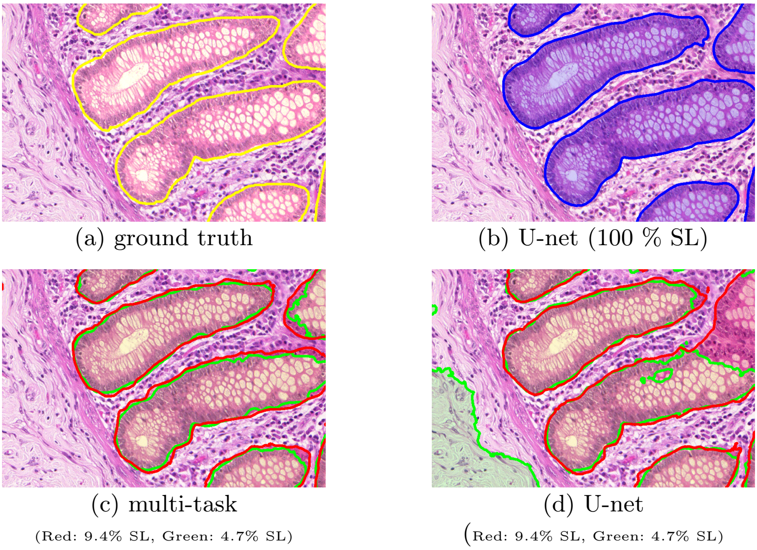
We apply the approach to the segmentation of tissues in histology images. In this experiment, we use the GlaS challenge dataset [37] that consists of 165 Hematoxylin and eosin (H&E) stained images. The dataset is split into three parts, with 85 images for training, and 60 for offsite test and 20 images for onsite test (we will call the latter two sets Test part A and Test part B respectively in the following).

Apart from the SL available from the dataset, we create a set of a weak labels for the detection task and separation task. These weak labels together with a part of the strong labels are used for training the multi-task U-net.

Table 3. Average dice score for segmentation of gland. Results of two sets of methods, weakly supervised (WS) and strongly supervised (SS) are displayed. Our method uses both SL and WL. The ratio of strong labels (SL) is increased from 2.4% to 100%, and the scores of the methods are reported here for two parts A and B of the test sets, as split in [37].

SL Ratio			2.4%	4.7%	9.4%	100%
Test Part A	WS	Ours	**0.866**	**0.889**	**0.915**	**0.921**
		Single task	0.700	0.749	0.840	0.921
		PL	0.799	0.812	0.820	
	SS	MDUnet				0.920
Test Part B	WS	Ours	**0.751**	**0.872**	**0.904**	**0.910**
		Single task	0.658	0.766	0.824	0.908
		PL	0.773	0.770	0.782	
	SS	MDUnet				0.871

In this experiment, we test the algorithm on different ratios of SL, and compare it with the baseline U-net (single task), PL approach (where PL are generated in the same way as the ones for the SEM dataset), and a fully supervised approach called Multi-scale Densely Connected U-Net (MDUnet) [45]. The results on two sets of test data are reported in Table 3. As the SL ratios increase from 2.4% to 9.4%, an improvement of performance of the multi-task U-net is gained. When it reaches 9.4% SL, the multi-task framework achieves comparable score with the fully supervised version, and outperforms the PL approach by a significant margin. We emphasize that the 9.4% SL and WL can be obtained several times faster than the 100% SL used for fully supervised learning. Example of segmentation results are displayed in Figure 7.



(a) ground truth

(b) U-net (100 % SL)

(c) multi-task

(Red: 9.4% SL, Green: 4.7% SL)

(d) U-net

(Red: 9.4% SL, Green: 4.7% SL)

Fig. 7. Segmentation results on the gland dataset (best view in color). The ground truth and the results. For (c) and (d), Red contour denotes the results from 9.4% strong labels; Green contour denotes results from 4.7% strong labels.

5 Conclusion

In this paper, we develop a multi-task learning framework for microscopy image segmentation, which relaxes the requirement for numerous and accurate annotations to train the network. It is therefore suitable for segmentation problem with a dense population of object instances. The model separates the segmentation problem into three smaller tasks. One of them is dedicated to the instance detection and therefore does not need exact boundary information. This gives potential flexibility as one could concentrate on the classification and rough location of the instances during data collection. The second one focuses on the separation of objects sharing a common boundary. The final task aims at extracting pixel-wise boundary information. Thanks to the information shared within the multi-task learning, this accurate segmentation can be obtained using very few annotated data.

Our model learns directly the statistics of weal labels WL as auxiliary tasks, and no further processing steps are needed before training the network. For the partial masks that ignore boundary pixels, the annotation can also be done when the boundaries of object are hard to detect. As a small amount of strong labels SL is needed and the collection of WL can be fast and cheap, the proposed framework is potentially effective for applications with growing datasets. The weakly annotated set for detection purpose could be augmented if necessary and the new images could easily be incorporated into our end-to-end framework. In the future, we could like to extend the proposed approach for solving 3D segmentation problems in biomedical images where labelling a single 3D image needs much more manual work.

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence **39**(12), 2481–2495 (2017)
2. Bajcsy, P., Feldman, S., Majurski, M., Snyder, K., Brady, M.: Approaches to training multiclass semantic image segmentation of damage in concrete. Journal of Microscopy (2020)
3. Bearman, A., Russakovsky, O., Ferrari, V., Fei-Fei, L.: What's the point: Semantic segmentation with point supervision. In: European conference on computer vision. pp. 549–565. Springer (2016)
4. Bischke, B., Helber, P., Folz, J., Borth, D., Dengel, A.: Multi-task learning for segmentation of building footprints with deep neural networks. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 1480–1484. IEEE (2019)
5. Boykov, Y., Funka-Lea, G.: Graph cuts and efficient nd image segmentation. International journal of computer vision **70**(2), 109–131 (2006)
6. Buchholz, T.O., Prakash, M., Krull, A., Jug, F.: Denoiseg: Joint denoising and segmentation. arXiv preprint arXiv:2005.02987 (2020)
7. Caselles, V., Kimmel, R., Sapiro, G.: Geodesic active contours. International journal of computer vision **22**(1), 61–79 (1997)
8. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062 (2014)
9. 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 transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2018)
10. Ciofolo, C., Barillot, C.: Atlas-based segmentation of 3d cerebral structures with competitive level sets and fuzzy control. Medical image analysis **13**(3), 456–470 (2009)
11. Ezhov, M., Zakirov, A., Gusarev, M.: Coarse-to-fine volumetric segmentation of teeth in cone-beam ct. arXiv preprint arXiv:1810.10293 (2018)
12. Ghosh, A., Ehrlich, M., Shah, S., Davis, L., Chellappa, R.: Stacked u-nets for ground material segmentation in remote sensing imagery. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 257–261 (2018)
13. Guerrero-Pena, F.A., Fernandez, P.D.M., Ren, T.I., Yui, M., Rothenberg, E., Cunha, A.: Multiclass weighted loss for instance segmentation of cluttered cells. In: 2018 25th IEEE International Conference on Image Processing (ICIP). pp. 2451–2455. IEEE (2018)
14. Heimann, T., Meinzer, H.P.: Statistical shape models for 3d medical image segmentation: a review. Medical image analysis **13**(4), 543–563 (2009)
15. Hirsch, P., Kainmueller, D.: An auxiliary task for learning nuclei segmentation in 3d microscopy images. arXiv preprint arXiv:2002.02857 (2020)
16. Hong, S., Noh, H., Han, B.: Decoupled deep neural network for semi-supervised semantic segmentation. In: Advances in neural information processing systems. pp. 1495–1503 (2015)
17. Huang, Z., Wang, X., Wang, J., Liu, W., Wang, J.: Weakly-supervised semantic segmentation network with deep seeded region growing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7014–7023 (2018)

18. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
19. Jing, L., Chen, Y., Tian, Y.: Coarse-to-fine semantic segmentation from image-level labels. arXiv preprint arXiv:1812.10885 (2018)
20. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7482–7491 (2018)
21. Khoreva, A., Benenson, R., Hosang, J., Hein, M., Schiele, B.: Simple does it: Weakly supervised instance and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 876–885 (2017)
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
23. Kolesnikov, A., Lampert, C.H.: Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In: European Conference on Computer Vision. pp. 695–711. Springer (2016)
24. Krasowski, N., Beier, T., Knott, G., Köthe, U., Hamprecht, F.A., Kreshuk, A.: Neuron segmentation with high-level biological priors. IEEE transactions on medical imaging **37**(4), 829–839 (2017)
25. Lee, J., Kim, E., Lee, S., Lee, J., Yoon, S.: Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. arXiv preprint arXiv:1902.10421 (2019)
26. Lin, D., Dai, J., Jia, J., He, K., Sun, J.: Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3159–3167 (2016)
27. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis **42**, 60–88 (2017)
28. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
29. MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. vol. 1, pp. 281–297. Oakland, CA, USA (1967)
30. Mlynarski, P., Delingette, H., Criminisi, A., Ayache, N.: Deep learning with mixed supervision for brain tumor segmentation. arXiv preprint arXiv:1812.04571 (2018)
31. Papandreou, G., Chen, L.C., Murphy, K.P., Yuille, A.L.: Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: Proceedings of the IEEE international conference on computer vision. pp. 1742–1750 (2015)
32. Playout, C., Duval, R., Cheriet, F.: A novel weakly supervised multitask architecture for retinal lesions segmentation on fundus images. IEEE transactions on medical imaging (2019)
33. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
34. Rother, C., Kolmogorov, V., Blake, A.: Grabcut: Interactive foreground extraction using iterated graph cuts. In: ACM transactions on graphics (TOG). vol. 23, pp. 309–314. ACM (2004)
35. Shah, M.P., Merchant, S., Awate, S.P.: Ms-net: Mixed-supervision fully-convolutional networks for full-resolution segmentation. In: International Conference

on Medical Image Computing and Computer-Assisted Intervention. pp. 379–387. Springer (2018)

36. Shin, S.Y., Lee, S., Yun, I.D., Kim, S.M., Lee, K.M.: Joint weakly and semi-supervised deep learning for localization and classification of masses in breast ultrasound images. IEEE transactions on medical imaging **38**(3), 762–774 (2019)
37. Sirinukunwattana, K., Pluim, J.P., Chen, H., Qi, X., Heng, P.A., Guo, Y.B., Wang, L.Y., Matuszewski, B.J., Bruni, E., Sanchez, U., et al.: Gland segmentation in colon histology images: The glas challenge contest. Medical image analysis **35**, 489–502 (2017)
38. Sun, F., Li, W.: Saliency guided deep network for weakly-supervised image segmentation. Pattern Recognition Letters **120**, 62–68 (2019)
39. Sun, T., Chen, Z., Yang, W., Wang, Y.: Stacked u-nets with multi-output for road extraction. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 187–1874. IEEE (2018)
40. Tsutsui, S., Kerola, T., Saito, S., Crandall, D.J.: Minimizing supervision for free-space segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 988–997 (2018)
41. Wang, C., MacGillivray, T., Macnaught, G., Yang, G., Newby, D.: A two-stage 3d unet framework for multi-class segmentation on full resolution image. arXiv preprint arXiv:1804.04341 (2018)
42. Wang, X., Xiao, T., Jiang, Y., Shao, S., Sun, J., Shen, C.: Repulsion loss: Detecting pedestrians in a crowd. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7774–7783 (2018)
43. Wei, Y., Liang, X., Chen, Y., Shen, X., Cheng, M.M., Feng, J., Zhao, Y., Yan, S.: Stc: A simple to complex framework for weakly-supervised semantic segmentation. IEEE transactions on pattern analysis and machine intelligence **39**(11), 2314–2320 (2017)
44. Wei, Y., Xiao, H., Shi, H., Jie, Z., Feng, J., Huang, T.S.: Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7268–7277 (2018)
45. Zhang, J., Jin, Y., Xu, J., Xu, X., Zhang, Y.: Mdu-net: Multi-scale densely connected u-net for biomedical image segmentation. arXiv preprint arXiv:1812.00352 (2018)
46. Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H.: Conditional random fields as recurrent neural networks. In: Proceedings of the IEEE international conference on computer vision. pp. 1529–1537 (2015)
47. Zhou, J., Luo, L.Y., Dou, Q., Chen, H., Chen, C., Li, G.J., Jiang, Z.F., Heng, P.A.: Weakly supervised 3d deep learning for breast cancer classification and localization of the lesions in mr images. Journal of Magnetic Resonance Imaging (2019)
48. Zhou, S., Nie, D., Adeli, E., Yin, J., Lian, J., Shen, D.: High-resolution encoder–decoder networks for low-contrast medical image segmentation. IEEE Transactions on Image Processing **29**, 461–475 (2019)