Learning to segment microscopy images with lazy labels

Rihuan Ke, Aurélie Bugeau, Nicolas Papadakis, Peter Schuetz, Carola-Bibiane Schönlieb

A multi-task U-net for segmentation with lazy labels

Rihuan Ke
University of Cambridge
rk621@cam.ac.uk

Aurélie Bugeau
University of Bordeaux
aurelie.bugeau@labri.fr

Nicolas Papadakis
CNRS, IMB
nicolas.papadakis@math.u-bordeaux.fr

Peter Schuetz
Unilever R&D Colworth
peter.schuetz@unilever.com

Carola-Bibiane Schönlieb
University of Cambridge
cbs31@cam.ac.uk

Abstract

The need for labour intensive pixel-wise annotation is a major limitation of many fully supervised learning methods for image segmentation. In this paper, we propose a deep convolutional neural network for multi-class segmentation that circumvents this problem by being trainable on coarse data 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 then stratified into three connected tasks: rough detection of class instances, separation of wrongly connected objects without a clear boundary, and pixel-wise segmentation to find the accurate boundaries of each object. These problems are integrated into a multi-task learning framework and the model is trained end-to-end in a semi-supervised fashion. The method is applied on a dataset of food microscopy images. We show that the model gives accurate segmentation results even if exact boundary labels are missing for a majority of the annotated data. This allows more flexibility and efficiency for training deep neural networks that are data hungry in a practical setting where manual annotation is expensive, by collecting more lazy (rough) annotations than precisely segmented images.

1 Introduction

Image segmentation has been an active research field in the past decades. Deep learning approaches play an increasingly important role and has become state-of-the-art in various segmentation tasks [17, 21, 41, 13, 26]. Though fully supervised segmentation neural networks have shown great success, one of their most challenging issues is the need for pixel-level annotations. Obtaining such annotations usually requires a great and expensive amount of manual work.

In this paper, we propose a multi-class and multi-instance segmentation approach that we split into three relevant tasks: detection, separation and segmentation. Doing so, we obtain a semi-supervised learning approach with that is trained with so-called "lazy" labels, that is a lot of coarse annotations of class instances together with only a few pixel-wise annotated images that can be obtained from the coarse labels in a semi-automated way. In the following, we will refer to weak (resp. strong) annotations for coarse (resp. accurate) labels. Cf. Figure 1 for a demonstration of our proposed segmentation algorithm.

Preprint. Under review.

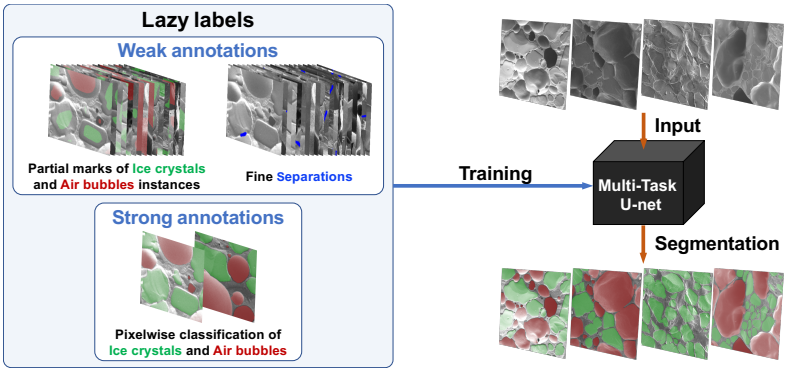

Figure 1: Multi-task learning for image segmentation with lazy labels, demonstrated for the pixel-wise classification of Scanning Electron Microscopy (SEM) images of ice cream into air bubbles (green), ice crystals (red) and background. Most of the training data are weak annotations containing (i) partial 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 top right several SEM images are displayed. Their corresponding output, obtained with the trained network, are shown at the bottom right.

Task 1 of our multi-class segmentation approach is instance detection. It detects and classifies each object and roughly determines its region through an under-segmentation mask. Instance counting can be obtained as a by-product of this task. At this stage, exact labels for the whole object or its boundary are not necessary. Instead, this task is trained on weakly annotated images in which a rough region inside each object is marked, cf. the most top left part of Figure 1.
Task 2 focuses on the separation of instances that are connected without a clear boundary dividing them. Separating objects sharing a common boundary is a well known challenge when segmenting images containing a dense population of instances such as cells [14, 34] or people in a crowd [43]. Also for this task we rely on weak annotations to reduce the burden of manual annotations: touching interfaces are specified with rough scribbles, cf. top left part of Figure 1.
Task 3 finally tackles the pixel-wise classification of the instances. It requires strong annotations that are accurate up to the boundaries of the objects. Thanks to the information brought by the 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. Our model is based on the U-net [34] with all three tasks sharing the same contracting path, and we introduce a new multi-task block for the expansive path. The network has three outputs and is fed with a combination of weak and strong labels described above. Since weakly and strongly annotated training data is shared between the tasks, part of the annotations are missing, especially for task 3. To accommodate for this we introduce a weighted loss function over the samples. Accurate segmentation labels for training are usually not easy to obtain. With our approach we demonstrate that exact labels for the whole training set are not needed for good segmentation learning performance.

We demonstrate the performance of our approach on the segmentation of food microstructures in Scanning Electron Microscopy (SEM) images of ice cream. 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 food 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 top 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.

In summary, our contributions are as follows.

1. We present a semi-supervised learning strategy for segmentation with lazy labels, made of coarse instance detections, sparse scribbles, and a small set of accurate segmentation masks.
2. We develop a multi-task learning framework that integrates the instance detection, separation and segmentation within a deep neural network. The network is trained end-to-end.
3. We propose a new multi-task building block for the purpose of joint training of the detection/separation/segmentation tasks and use it at multiple resolutions.

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 being developed, including traditional unsupervised methods like k-means that splits the image into homogeneous regions according to image low level features, curve evolution based methods like snakes [6], graph-cut based methods like Grabcut [35], just to name a few. Interactive approaches such as snakes or Grabcut enable getting involved users' knowledge by means of initializing regions or putting constraints on the segmentation results.

Deep convolutional neural network (DCNN) approaches have been developed for segmenting complex images, especially in the semantic setting. In particular, fully convolutional networks (FCN) [27] 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 [7] 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].

Other types of FCNs are deep encoder-decoder networks [2, 34]. They have multiple up-sampling layers for better localizing boundary details. One of the most well-known models is the U-net [34]. 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, and an expansive path that has an equal amount of up-sampling layers. The higher resolution feature maps on the contracting path are merged with up-sampled layers via long skip connections to recover boundary information after down-sampling.

Weakly/semi-supervised segmentation learning. Motivated by the heavy cost of pixel-level annotation needed for fully supervised learning, there has been a growing interest in weakly supervised learning with coarse annotations for semantic segmentation in the last years. Common weak annotations include image-level labels [17, 29, 24], bounding boxes[21], scribbles [25] and points [4].
Most 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 [45, 21, 24, 44, 41]. The segmentation masks can be improved recursively, which involves several rounds of training of the segmentation network [45, 19, 12].

Other techniques do not use any explicit pseudo segmentation masks as training data, but instead compute a composite loss from other guiding principles. For instance, the SEC method [23] combines localization cues (Seed), classification coherence (Expand) and Constrain-to-boundary with a fully-connected CRF. Special modules could be designed to further exploit the information of sources, such as the Deep Seeded Region Growing [17] and saliency seeded region growing [39].

Semi-supervised deep learning provides an alternative way to reduce the pixel-level annotation burden [3, 30]. These methods require only a small amount of strongly annotated data and a large set of unlabelled data. Several training methods that combine weak annotations with a limited set of strong annotations have already been explored [29, 44, 24]. The pixel-wise labels can be integrated through an additional loss function, and used along with their semi-supervised counterparts. Better weak supervision performance can be obtained by feeding the network with learned object localization maps [44, 24]. However, the networks are trained with image level labels from large scale datasets like ImageNet [10] or PASCAL VOC 2012 [11].

The weakly and semi-supervised deep learning methods are also explored for object localization or segmentation on relatively smaller datasets, especially in medical imaging. For these methods, the weak annotations can be of image level [28, 33, 47, 38] or in forms of bounding box [37].

Multi-task learning. Multi-task learning algorithms consider several related tasks to improve the overall performance, taking benefits from the underlying common information that may be ignored by a single task learning. In deep neural based multi-task models, the common information is conveyed by soft or hard parameter sharing [36]. Various multi-task deep learning methods have been developed for segmentation, for example, the stacked U-net for extracting roads from satellite imagery [40], the two stage 3D U-net framework for 3D CT or MR data segmentation [42], encoder decoder networks for depth regression, semantic and instance segmentation [20], and for building footprint segmentation [5]. Compared to these works, our method handles the tasks with a multi-task block at different feature resolutions and is designed upon both strong and weak notations.

Our work is more closely related to the multi-task learning methods in [32, 33] for retinal lesions segmentation and [28] for brain tumour segmentation. Three tasks are considered in [33]: red lesion segmentation, bright lesions segmentation and image level lesion detection. The proposed encoder-decoder architecture has one branch in the encoder part and two branches into decoder part specified for the red/bright lesion segmentation and the image level classification respectively. The network is weakly supervised as only image level labels are used in one of the training phases. In [28], another U-net architecture is used for jointly segmenting and classifying brain tumours. Built upon a standard U-net structure for the segmentation task, an additional branch takes the last but one layer of the U-net as input followed by a mean pooling layer, and outputs a score for the classification task.

Our network structure, similar to these architectures, shares an encoder part for all tasks. The learning is also supervised by a mixture of strong annotations and weak annotations. However, the proposed method does not include supervision from image level annotations and is more specialized on distinguishing the different object instances and clarifying their boundaries in every single image.

Segmentation of SEM images of ice cream. As previously mentioned, multi-task segmentation methods are designed with respect to the tackled application. We here deal with a dense population of instances that arise in SEM images of ice cream. Most works on ice cream analysis are done by the food quality community. Standard techniques first require manual identification of air bubbles and ice crystals [16, 9], which is a major limitation in practice. Rough segmentations are then obtained using image processing techniques: anisotropic diffusion in [31], watershed in [1] or total variation in [15]. The proposed method is a cheaper and more accurate alternative to such processes.

3 Multi-task learning framework

Let $\mathcal{I}$ be the set of images and let $\mathcal{T}_k \subset \mathcal{I}$, $k = 1, 2, 3$, be the subset of images labelled for task k. The set $\mathcal{I}$ is randomly drawn from an unknown distribution. As we collect a different amount of annotations for each task, the number of annotated images $|\mathcal{T}_k|$ may not be the same for different k.

For a given image $I \in \mathcal{I}$, we denote as $\boldsymbol{s}^{(1)}$ the instance detection mask, $\boldsymbol{s}^{(2)}$ the map of interfaces shared by touching objects, and $\boldsymbol{s}^{(3)}$ the full segmentation mask. In the multi-task learning setting, one aims at approximating the conditional probability $p\Big(\boldsymbol{s}^{(1)}\Big|I\Big)$, $p\Big(\boldsymbol{s}^{(2)}\Big|I\Big)$, and $p\Big(\boldsymbol{s}^{(3)}\Big|I\Big)$. Let's consider the probabilities given by the models $p\Big(\boldsymbol{s}^{(k)}\Big|I;\boldsymbol{\theta}\Big)$ $(k = 1, 2, 3)$ parametrized by $\boldsymbol{\theta}$ which is determined such that the models match the desired probability distributions. The set of samples in $\mathcal{T}_3$ for segmentation being small, the computation of an accurate approximation of the true probability distribution $p\Big(\boldsymbol{s}^{(3)}\Big|I\Big)$ is a challenging issue. As a consequence, we do not optimize $\boldsymbol{\theta}$ for each individual task, but instead consider a joint probability $p\Big(\Big(\boldsymbol{s}^{(1)}, \boldsymbol{s}^{(2)}, \boldsymbol{s}^{(3)}\Big) \mid I\Big)$. Assuming that $\boldsymbol{s}^{(1)}$ (rough under-segmented instance detection) and $\boldsymbol{s}^{(2)}$ (rough 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 s

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

The set $|\mathcal{T}_3|$ 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 $\Big\{p\Big(\boldsymbol{s}^{(3)} \mid \boldsymbol{s}^{(1)}, \boldsymbol{s}^{(2)}, I\Big)\Big\}_{I\in\mathcal{I}}$ can be approximated correctly

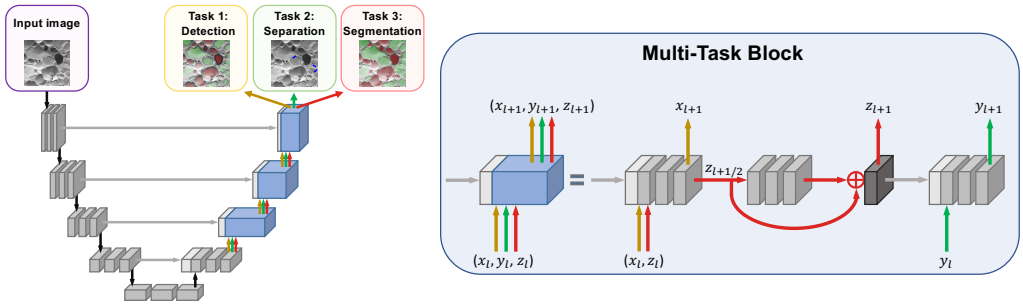

Figure 2: Architecture of the multi-task U-net. The left part of the network is a contracting path similar to the standard U-net. For multi-task learning, we construct several expansive paths with specific multi-task blocks. At each resolution, task 1 (Detection in yellow) and task 3 (Segmentation in 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.

by $p\Big(\boldsymbol{s}^{(3)} \mid \boldsymbol{s}^{(1)}, \boldsymbol{s}^{(2)}, I; \boldsymbol{\theta}\Big)$ even if $\boldsymbol{\theta}$ is computed without $\boldsymbol{s}^{(3)}$ specified for $\mathcal{I} \backslash \mathcal{T}_3$, then

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

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

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

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

Network architecture. We follow a convolutional encoder-decoder network structure for the 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

$$\boldsymbol{x}_{l+1} = F_{W_l}(\boldsymbol{x}_l, \boldsymbol{c}_l), \qquad \boldsymbol{z}_{l+\frac{1}{2}} = F_{W_l}(\boldsymbol{z}_l, \boldsymbol{c}_l), \qquad \boldsymbol{z}_{l+1} = \boldsymbol{z}_{l+\frac{1}{2}} + F_{W_{l+\frac{1}{2}}}(\boldsymbol{z}_{l+\frac{1}{2}}), \tag{4}$$

in which $W_l, W_{l+1/2} \in \boldsymbol{\theta}$ are parameters of the network. For task 2 the output at l^{th} block $\boldsymbol{y}_{l+1}$ is computed by $\boldsymbol{y}_l = G_{\tilde{W}_l}(\boldsymbol{z}_{l+1}, \boldsymbol{y}_l)$ 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. Let the output of the approximation models be denoted respectively by $h^{(1)}_{\boldsymbol{\theta}}(I)$, $h^{(2)}_{\boldsymbol{\theta}}(I)$, and $h^{(3)}_{\boldsymbol{\theta}}(I)$, with $\left[h^{(k)}_{\boldsymbol{\theta}}(I)\right]_{i,c}$ the estimated probability of pixel i being in class c of task k. For the label $\boldsymbol{s}^{(k)}$ of I, the log likelihood function for each task is decomposed into

$$\log p\Big(\boldsymbol{s}^{(k)} \mid I; \boldsymbol{\theta}\Big) = \sum_i \sum_{c \in C_k} s^{(k)}_{i,c} \log \left[h^{(k)}_{\boldsymbol{\theta}}(I)\right]_{i,c}, \quad k = 1, 2, 3, \tag{5}$$

in which $s^{(k)}_{i,c}$ 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 ice cream images, we have three classes (air bubbles, ice crystals and background), so $C_1, C_3 = \{1, 2, 3\}$. For the separation task, there are only two classes (touching instance interfaces and otherwise): $C_2 = \{1, 2\}$. According to Equation (3), the network is trained by minimizing the weighted cross entropy loss:

$$L\left(\boldsymbol{\theta}, \boldsymbol{s}^{(1)}, \boldsymbol{s}^{(2)}, \boldsymbol{s}^{(3)}\right) = -\sum_{I \in \mathcal{I}} \sum_{k=1}^{3} \alpha_k \mathbb{1}_{\mathcal{T}_k}(I) \log p\Big(\boldsymbol{s}^{(k)} \mid I; \boldsymbol{\theta}\Big), \tag{6}$$

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

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 [34], each small gray 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 gray 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.

4 Lazy labels generation

We now explain our strategy for generating all the lazy annotations that are used for training the multi-task U-net. We apply our method to a data set of ice cream SEM images but any other similar dataset 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 as 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 labeling 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 [6]. Unfortunately, such method fails to capture sharp corners and the contour evolution tends to ignore weak boundaries. The annotation then requires frequent and time consuming user interaction. Instead, we use Grabcut [35] 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.

Algorithm 1 Multi-task learning for segmentation with lazy labels

procedure LAZY LABELS($\mathcal{I}$) ▷ Choose $\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3 \subset \mathcal{I}$, $|\mathcal{T}_3|$ can be relatively small.
Input: set of images $\mathcal{I}$
  Select inner regions for each object in $\mathcal{T}_1$ for the detection task
  Indicate scribbles on images of $\mathcal{T}_2$ for the separation task
  Generate a few pixel-wise labels $\mathcal{T}_3$ from $\mathcal{T}_1$ using interactive segmentation tools (e.g., Grabcut)
  return $\mathcal{T}_1$, $\mathcal{T}_2$, $\mathcal{T}_3$ and the labels

procedure MULTI-TASK U-NET TRAINING($\mathcal{T}_k$, $\boldsymbol{s}_i^{(k)}$, α_k,r) ▷ $\boldsymbol{s}_i^{(k)}$ denote the labels
Input: labelled sets $\mathcal{T}_k$, loss function weights α_k for $k = 1, 2, 3$, Adam parameters r, mini-batch size m.
  Set the 1^{st} and 2^{nd} momentum vectors $\boldsymbol{m}$, $\boldsymbol{v}$ as zeros.
  Initialize the multi-task U-net parameter $\boldsymbol{\theta}$.
  while Termination criterion is not satisfied **do**
    Obtain a mini-batch $(I_1, \boldsymbol{s}_1^{(k)}), \cdots, (I_m, \boldsymbol{s}_m^{(k)})$. ▷ $\boldsymbol{s}_i^{(k)}$ can be unknown for some (i, k).
    Compute the gradient $\boldsymbol{g} \leftarrow -\nabla_{\boldsymbol{\theta}} \frac{1}{m} \sum_{k=1}^{3} \sum_{i=1}^{m} \alpha_k \mathbb{1}_{\mathcal{T}_k}(I_i) \log p\left(\boldsymbol{s}_i^{(k)} \mid I_i; \boldsymbol{\theta}\right)$
    $(\boldsymbol{\theta}, \boldsymbol{m}, \boldsymbol{v}) \leftarrow \text{Adam}_r(\boldsymbol{\theta}, \boldsymbol{m}, \boldsymbol{v}, \boldsymbol{g})$ ▷ $\text{Adam}_r(\cdot)$ is an Adam iteration
  return $\boldsymbol{\theta}$

5 Experiments

The training set contains 20 images of ice cream in a variety of sample conditions. Among these images, only 2 are manually labelled for task 3, while 15 annotations are available for task 1 and 20 for task 2. The validation-set includes 6 images annotated for all the tasks.

With the small set of lazy labelled images, data augmentation is crucial 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 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. The overall method is summarized with the 2 procedures presented in Algorithm 1.

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.

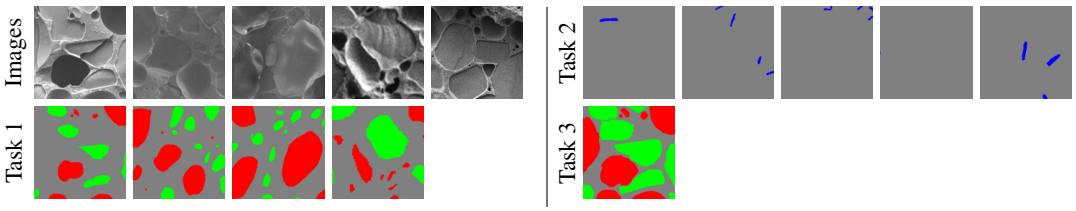

Figure 3: Example of annotated images. Some of the annotations are missing because not all images are labelled for task 1 and task 2. The marks in red are for air bubbles and the ones in green are for ice crystal instances. The blue curves on the third row are labels for interfaces of touching objects.

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 images, and average results are shown in Table 1. In the table, the dice score for a class c is defined as $d_c = 2\sum_i x_{i,c} y_{i,c} / \left(\sum_i x_{i,c} + \sum_i y_{i,c}\right)$where $\boldsymbol{x}$ is the computed segmentation mask and $\boldsymbol{y}$ 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 give 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

Table 1: Dice scores of segmentation results on 12 test images.

The models	air bubbles	ice crystals	Overall
U-net on weak annotations	0.73	0.71	0.72
U-net on strong annotations	0.84	0.79	0.82
pseudo label supervision approach [21]	0.94	0.91	0.92
Multi-task U-net	0.95	0.93	0.94

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 U-net trained on strong annotations only is scored at 0.82, which is still significantly lower than the 0.94 obtained by the multi-task U-net.

We also compare our multi-task U-net results with a weakly supervised learning strategy [21] that makes use of pseudo labels (PL). Specifically, the pseudo segmentation masks are created from the weak annotations and are used to feed a segmentation network. In our setting we use the rough inner regions as weak annotations instead of bounding boxes. Grabcut has been used to generate and improve the segmentation masks [21, 19] for training purpose. We adopt it in the segmentation mark preparation without user correction, and segmentation network for this test is the baseline single task U-net.

Note that the approach [21] is originally designed for weakly supervised learning settings, but it can be easily adapted for semi-supervised learning by including a few accurate segmentation masks in the pseudo label set and then follows a standard fully supervised training routine with a segmentation network. As shown in Table 1, the dice score of PL is slightly lower than our multi-task U-net. A significant limitation of the PL method is that its performance heavily relies on the tools used for pseudo segmentation mask generations. Having common errors, instead of random ones, on the object boundaries in the training data, the segmentation network of PL also learns to have those patterns in the prediction. The images in the left part of Figure 4 show that the predicted label of an object tends to merge with some background pixels when there are edges of another object nearby. Similar errors from the GrabCut are illustrated in the supplementary material.

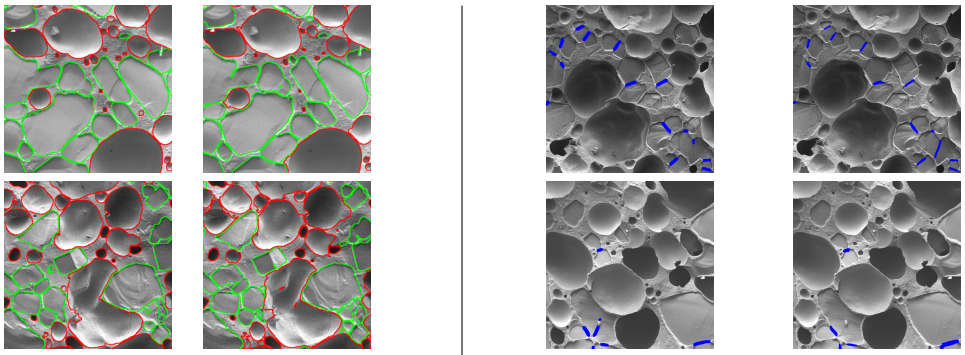

Figure 4: Segmentation and separation results (best viewed in color). Left: the computed contours are shown in red for air bubbles and green for ice crystals. While multi-task U-net and PL results are globally close, PL misclassifies the background near object boundaries. Right: Examples of separation by the multi-task U-net and the ground truth.

Besides the number of pixels that are correctly classified, the separation of touching instances is also of interest. The dice scores in Table 1 do not reflect whether the close instances are well separated. We study the learning performance of multi-task U-net 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 4.

6 Conclusion

In this paper, we develop a multi-task learning framework for image segmentation problems, with a focus on dense micro-structures of food. The model separates the segmentation problem into three smaller tasks. One of them is dedicated to the instance detection and therefore do 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.

As in many practical applications it is difficult to determine exactly how much labeled data is necessary beforehand, further data collection may be considered if the expected performance is not reached. In particular, the relatively cheap weakly annotated set for detection purpose could be augmented if necessary for characterizing a wider range of imaging conditions or situations. The increase of data sets could be easily incorporated into our framework as it does not follow a detection then segmentation pipeline, and it is trained end-to-end.

Acknowledgments

This work was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 777826. RK acknowledges supports from Unilever & EPSRC IAA Partnership Development Award. CBS acknowledges support from the Leverhulme Trust project on Breaking the non-convexity barrier, the Philip Leverhulme Prize, the EPSRC grant Nr. EP/M00483X/1, the EPSRC Centre Nr. EP/N014588/1, the RISE projects CHiPS, the Cantab Capital Institute for the Mathematics of Information and the Alan Turing Institute. We gratefully acknowledge the support of NVIDIA Corporation with the donation of a Quadro P6000 and a Titan Xp GPUs used for this research.

References

[1] G Alvarez, Dennis Cantre, Pieter Verboven, FT Ndoye, M Warren, WR Hartel, and Bart Nicolai. "X-ray micro-tomography to quantify frozen ice cream structure". In: *24ième Congrès International du Froid ICR 2015*. 2015, 8–p.

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation". In: *IEEE transactions on pattern analysis and machine intelligence* 39.12 (2017), pp. 2481–2495.

[3] Christoph Baur, Shadi Albarqouni, and Nassir Navab. "Semi-supervised deep learning for fully convolutional networks". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pp. 311–319.

[4] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. "What's the point: Semantic segmentation with point supervision". In: *European conference on computer vision*. Springer. 2016, pp. 549–565.

[5] Benjamin Bischke, Patrick Helber, Joachim Folz, Damian Borth, and Andreas Dengel. "Multi-task learning for segmentation of building footprints with deep neural networks". In: *arXiv preprint arXiv:1709.05932* (2017).

[6] Vicent Caselles, Ron Kimmel, and Guillermo Sapiro. "Geodesic active contours". In: *International journal of computer vision* 22.1 (1997), pp. 61–79.

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs". In: *IEEE transactions on pattern analysis and machine intelligence* 40.4 (2018), pp. 834–848.

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. "Semantic image segmentation with deep convolutional nets and fully connected crfs". In: *arXiv preprint arXiv:1412.7062* (2014).

[9] G van Dalen, J Bent, E Guo, G Zeng, P Rockett, and PD Lee. "A comparative study of desktop-and synchrotron radiation-based X-ray microtomography analysing air bubbles and ice crystals in ice cream". In: ().

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[11] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. "The pascal visual object classes challenge: A retrospective". In: *International journal of computer vision* 111.1 (2015), pp. 98–136.

[12] Matvey Ezhov, Adel Zakirov, and Maxim Gusarev. "Coarse-to-fine volumetric segmentation of teeth in Cone-Beam CT". In: *arXiv preprint arXiv:1810.10293* (2018).

[13] Arthita Ghosh, Max Ehrlich, Sohil Shah, Larry Davis, and Rama Chellappa. "Stacked U-Nets for Ground Material Segmentation in Remote Sensing Imagery". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018, pp. 257–261.

[14] Fidel A Guerrero-Pena, Pedro D Marrero Fernandez, Tsang Ing Ren, Mary Yui, Ellen Rothenberg, and Alexandre Cunha. "Multiclass weighted loss for instance segmentation of cluttered cells". In: *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE. 2018, pp. 2451–2455.

[15] Enyu Guo, Daniil Kazantsev, Jingyi Mo, Julian Bent, Gerard Van Dalen, Peter Schuetz, Peter Rockett, David StJohn, and Peter D Lee. "Revealing the microstructural stability of a three-phase soft solid (ice cream) by 4D synchrotron X-ray tomography". In: *Journal of food engineering* 237 (2018), pp. 204–214.

[16] Enyu Guo, Guang Zeng, Daniil Kazantsev, Peter Rockett, Julian Bent, Mark Kirkland, Gerard Van Dalen, David S Eastwood, David StJohn, and Peter D Lee. "Synchrotron X-ray tomographic quantification of microstructural evolution in ice cream–a multi-phase soft solid". In: *Rsc Advances* 7.25 (2017), pp. 15561–15573.

[17] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. "Weakly-supervised semantic segmentation network with deep seeded region growing". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 7014–7023.

[18] Sergey Ioffe and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *arXiv preprint arXiv:1502.03167* (2015).

[19] Longlong Jing, Yucheng Chen, and Yingli Tian. "Coarse-to-fine Semantic Segmentation from Image-level Labels". In: *arXiv preprint arXiv:1812.10885* (2018).

[20] Alex Kendall, Yarin Gal, and Roberto Cipolla. "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*. 2018, pp. 7482–7491.

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

[22] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[23] Alexander Kolesnikov and Christoph H Lampert. "Seed, expand and constrain: Three principles for weakly-supervised image segmentation". In: *European Conference on Computer Vision*. Springer. 2016, pp. 695–711.

[24] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. "FickleNet: Weakly and Semi-supervised Semantic Image Segmentation using Stochastic Inference". In: *arXiv preprint arXiv:1902.10421* (2019).

[25] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. "Scribblesup: Scribble-supervised convolutional networks for semantic segmentation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 3159–3167.

[26] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. "A survey on deep learning in medical image analysis". In: *Medical image analysis* 42 (2017), pp. 60–88.

[27] Jonathan Long, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.

[28] Pawel Mlynarski, Hervé Delingette, Antonio Criminisi, and Nicholas Ayache. "Deep Learning with Mixed Supervision for Brain Tumor Segmentation". In: *arXiv preprint arXiv:1812.04571* (2018).

[29] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. "Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1742–1750.

[30] Christian S Perone and Julien Cohen-Adad. "Deep semi-supervised segmentation with weight-averaged consistency targets". In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 12–19.

[31] BR Pinzer, A Medebach, HJ Limbach, C Dubois, M Stampanoni, and M Schneebeli. "3D-characterization of three-phase systems using X-ray tomography: tracking the microstructural evolution in ice cream". In: *Soft Matter* 8.17 (2012), pp. 4584–4594.

[32] Clément Playout, Renaud Duval, and Farida Cheriet. "A Multitask Learning Architecture for Simultaneous Segmentation of Bright and Red Lesions in Fundus Images". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 101–108.

[33] Clément Playout, Renaud Duval, and Farida Cheriet. "A novel weakly supervised Multitask Architecture for Retinal Lesions Segmentation on Fundus Images". In: *IEEE transactions on medical imaging* (2019).

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

[35] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. "Grabcut: Interactive foreground extraction using iterated graph cuts". In: *ACM transactions on graphics (TOG)*. Vol. 23. 3. ACM. 2004, pp. 309–314.

[36] Sebastian Ruder. "An overview of multi-task learning in deep neural networks". In: *arXiv preprint arXiv:1706.05098* (2017).

[37] Meet P Shah, SN Merchant, and Suyash P Awate. "MS-Net: Mixed-supervision fully-convolutional networks for full-resolution segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 379–387.

[38] Seung Yeon Shin, Soochahn Lee, Il Dong Yun, Sun Mi Kim, and Kyoung Mu Lee. "Joint weakly and semi-supervised deep learning for localization and classification of masses in breast ultrasound images". In: *IEEE transactions on medical imaging* 38.3 (2019), pp. 762–774.

[39] Fengdong Sun and Wenhui Li. "Saliency guided deep network for weakly-supervised image segmentation". In: *Pattern Recognition Letters* 120 (2019), pp. 62–68.

[40] Tao Sun, Zehui Chen, Wenxiang Yang, and Yin Wang. "Stacked u-nets with multi-output for road extraction". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE. 2018, pp. 187–1874.

[41] Satoshi Tsutsui, Tommi Kerola, Shunta Saito, and David J Crandall. "Minimizing Supervision for Free-space Segmentation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018, pp. 988–997.

[42] Chengjia Wang, Tom MacGillivray, Gillian Macnaught, Guang Yang, and David Newby. "A two-stage 3D Unet framework for multi-class segmentation on full resolution image". In: *arXiv preprint arXiv:1804.04341* (2018).

[43] Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. "Repulsion loss: Detecting pedestrians in a crowd". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 7774–7783.

[44] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. "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*. 2018, pp. 7268–7277.

[45] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. "Stc: A simple to complex framework for weakly-supervised semantic segmentation". In: *IEEE transactions on pattern analysis and machine intelligence* 39.11 (2017), pp. 2314–2320.

[46] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. "Conditional random fields as recurrent neural networks". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1529–1537.

[47] Juan Zhou, Lu-Yang Luo, Qi Dou, Hao Chen, Cheng Chen, Gong-Jie Li, Ze-Fei Jiang, and Pheng-Ann Heng. “Weakly supervised 3D deep learning for breast cancer classification and localization of the lesions in MR images”. In: *Journal of Magnetic Resonance Imaging* (2019).