Sparse Object-level Supervision for Instance Segmentation with Pixel Embeddings

Adrian Wolny Qin Yu Constantin Pape Anna Kreshuk

European Molecular Biology Laboratory (EMBL), Heidelberg, Germany

Abstract

Most state-of-the-art instance segmentation methods have to be trained on densely annotated images. While difficult in general, this requirement is especially daunting for biomedical images, where domain expertise is often required for annotation and no large public data collections are available for pre-training. We propose to address the dense annotation bottleneck by introducing a proposal-free segmentation approach based on non-spatial embeddings, which exploits the structure of the learned embedding space to extract individual instances in a differentiable way. The segmentation loss can then be applied directly to instances and the overall pipeline can be trained in a fully- or weakly supervised manner. We consider the challenging case of positive-unlabeled supervision, where a novel self-supervised consistency loss is introduced for the unlabeled parts of the training data. We evaluate the proposed method on 2D and 3D segmentation problems in different microscopy modalities as well as on the Cityscapes and CVPPP instance segmentation benchmarks, achieving state-of-the-art results on the latter.

1. Introduction

Instance segmentation is one of the key problems addressed by computer vision. It is important for many application domains, from astronomy to scene understanding in robotics, forming the basis for the analysis of individual object appearance. Biological imaging provides a particularly large set of use cases for the instance segmentation task, with imaging modalities ranging from natural photographs for phenotyping to electron microscopy for detailed analysis of cellular ultrastructure. The segmentation task is often posed in crowded 3D environments or their 2D projections with multiple overlapping objects. Additional challenges – compared to segmentation in natural images – come from the lack of large, publicly accessible, annotated training datasets that could serve for general-purpose backbone training. Most microscopy segmentation networks are therefore trained from scratch, using annotations produced by domain experts in their limited time.

Over the recent years, several weakly supervised segmentation approaches have been introduced to lighten the necessary annotation burden. For natural images, image-level labels can serve as a surprisingly strong supervision thanks to the popular image classification datasets which include images of individual objects and can be used for pre-training [11]. There are no such collections in microscopy (see also Fig. 5 for a typical instance segmentation problem example where image-level labels would be of no help). Semi-supervised instance segmentation methods [4, 3, 9] can create pseudo-labels in the unlabeled parts of the dataset. However, these methods require (weak) annotation of *all* the objects in at least a subset of images – a major obstacle for microscopy datasets which often contain hundreds of tightly clustered objects, in 3D.

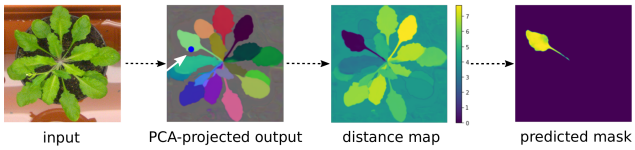

Figure 1. Differentiable instance selection for non-spatial embedding networks. First, we sample an anchor point randomly or guided by the groundtruth instances. Second, we compute a distance map in the embedding space from the anchor point to all image pixels. In the final step, a kernel function (Eq 3) transforms the distance map to the "soft" instance mask.

The aim of our contribution is to address the dense annotation bottleneck by proposing a different kind of weak supervision for the instance segmentation problem: we require mask annotations for a subset of instances in the image, leaving the rest of the pixels unlabeled. This "positive unlabeled" setting has been explored in image classification and semantic segmentation problems [34, 30], but - to the best of our knowledge - not for instance segmentation. Intrinsically, the instance segmentation problem is

Correspondence to: anna.kreshuk@embl.de
Code: github.com/kreshuklab/spoco

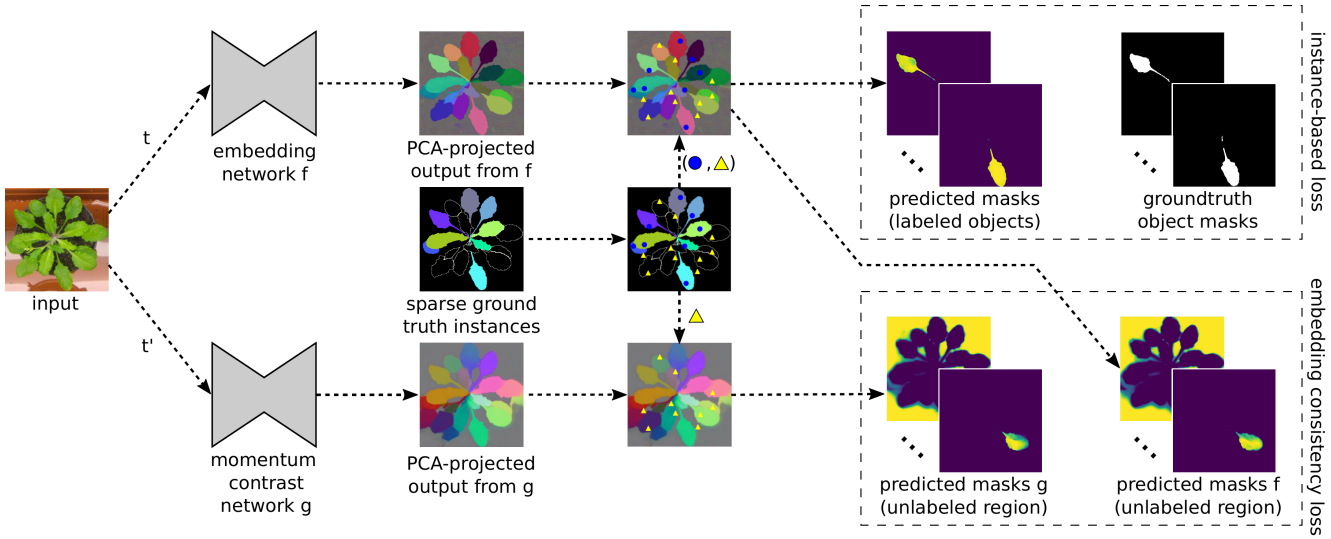

Figure 2. Overview of training procedure. Two augmented views of an input image are passed through two embedding networks $f(\cdot)$ and $g(\cdot)$ respectively. Anchor pixels inside labeled objects (blue dots) are sampled and their corresponding instances are extracted as shown in Fig. 1. Discrepancy between extracted objects and groundtruth objects is minimized by the instance-based loss. Another set of anchors (yellow triangles) is sampled exhaustively from the unlabeled region and for each anchor two instances are selected based on the outputs from $f(\cdot)$ and $g(\cdot)$. Discrepancy between instances is minimized using embedding consistency loss.

very well suited for positive unlabeled supervision: as we also show empirically (Appendix A.5), sampling a few objects in each image instead of labeling a few images densely exposes the network to a more varied training set with better generalization potential. This is particularly important for datasets with sub-domains in the raw data distribution, as it can ensure all sub-domains are sampled without increasing the annotation time. Furthermore, in crowded microscopy images which commonly contain hundreds of objects, and especially in 3D, dense annotation is significantly more difficult and time consuming than sparse annotation, for the same total number of objects annotated. The main obstacle for training an instance segmentation method on sparse object mask annotations lies in the assignment of pixels to instances that happens in a non-differentiable step which precludes the loss from providing supervision at the level of individual instances. To lift this restriction, we propose a differentiable instance selection step which allows us to incorporate any (differentiable) *instance-level* loss function into non-spatial pixel embedding network [6] training (Fig. 1). We show that with dense object mask annotations and thus full supervision, application of the loss at the single instance level consistently improves the segmentation accuracy of pixel embedding networks across a variety of datasets. For our main use case of weak positive unlabeled (PU) supervision, we propose to stabilize the training from sparse object masks by an additional instance-level consistency loss in the unlabeled areas of the images. The conceptually simple unlabeled consistency loss, inspired by [21, 46], does not require the estimation of class prior distributions or the propagation of pseudo-labels, ubiquitously present in PU and other weakly supervised segmentation approaches [47, 33]. In addition to training from scratch, our approach can deliver efficient domain adaptation using a few object masks in the target domain as supervision.

In summary, we address the instance segmentation task with a CNN that learns pixel embeddings and propose the first approach to enable training with weak positive unlabeled supervision, where only a subset of the object masks are annotated and no labels are given for the background. To this end, we introduce: (1) a differentiable instance selection step which allows to apply the loss directly to individual instances; (2) a consistency loss term that allows for instance-level training on unlabeled image regions, (3) a fast and scalable algorithm to convert the pixel embeddings into final instances, which partitions the metric graph derived from the embeddings. We evaluate our approach on natural images (CVPPP [37] , Cityscapes [14]) and microscopy datasets (2D and 3D, light and electron microscopy), reaching the state-of-the-art on CVPPP and consistently outperforming strong baselines for microscopy. On all datasets, the bulk of CNN performance improvement happens after just a fraction of training objects are annotated.

2. Related work

Proposal-based methods such as Mask R-CNN [22] are a popular choice for instance segmentation in natural images. These methods can be trained from weak bounding box labels [24, 32, 45, 39]. However, as they require a pre-trained backbone network and have difficulties segmenting complex non-convex shapes, they have not become

the go-to segmentation technique for microscopy imaging. There, instance segmentation methods commonly start from the semantic segmentation [43], followed by a (non-differentiable) post-processing [1, 17, 29, 40].

Semantic instance segmentation with embedding networks was introduced by [16, 6]. The embeddings of [6] have no explicit spatial or semantic component. [16] predicts a seediness score for each pixel in addition to the embedding vector. The main advantage of pixel embedding-based segmentation methods lies in their superior performance for overlapping objects and crowded environments, delivering state-of-the-art results in many benchmarks, including those for biological data [23]. Furthermore, they achieve a significant simplification of the pipeline: the same method can now be trained for intensity-based and for boundary-based segmentation. Our approach continues this line of work and employs non-spatial pixel embeddings.

Like the original proposal of [6], all modern embedding networks require fully segmented images for training and compute the loss for the whole image rather than for the individual instances. Even when the supervision annotations are weak, such as scribbles or saliency mask, they are commonly used to create full object proposals or pseudo-labels to allow the loss to be applied to the whole image [18, 47, 33]. Such methods exploit object priors learned by their components which have been pre-trained on large public datasets. At the moment, such datasets or pre-trained backbones are not available for microscopy images. Another popular approach to weak supervision is to replace mask annotations by bounding boxes [24] which are much faster to produce. Given a pre-trained backbone, bounding boxes can be reduced to single point annotations [27], but for training every object must be annotated, however weakly. The aim of our work is to lift this requirement and enable instance segmentation training with positive unlabeled supervision.

Positive unlabeled learning targets classification problems where negative labels are unavailable or unreliable [35, 2]. Three approaches are in common use: generation of negative pseudo-labels, biased learning with class label noise in unlabeled areas and class prior incorporation (see [2] for detailed review). PU learning has recently been extended to object detection [5] and semantic segmentation problems [31]. Our approach enables PU learning for instance segmentation problems via an instance-level consistency loss applied to the unlabeled areas.

The core of our approach consists of the differentiable single instance selection step performed during training. Here, we have drawn inspiration from [38], where the clustering bandwidth is learned in the network training which allows to optimize the intersection-over-union loss for each instance. Still, as the network also needs to be trained to predict a seed map of cluster centers for inference, this method cannot be trained on partially labeled images. Differentiable single instance selection has also been proposed by AdaptIS [44]. However, this method does not use a learned pixel embedding space and thus requires an additional sub-network to perform instance selection. Importantly, AdaptIS does not introduce PU training and relies on a pre-trained backbone network which is not readily available for microscopy images.

3. Methods

3.1. Full supervision

Given an image $I = \{I_1, ..., I_C\}$ composed of C objects (including background), N_k pixels in I_k, $N = \sum_{k=1}^{C} N_k$ pixels in the image and an embedding network $f\colon \mathbb{R}^3 \rightarrow \mathbb{R}^D$ which maps pixel i into a D-dimensional embedding vector $\boldsymbol{e}_i$, the discriminative loss [6] is defined by the *pull force* and the *push force* terms[1]:

$$L_{pull} = \frac{1}{C}\sum_{k=1}^{C}\frac{1}{N_k}\sum_{i=1}^{N_k}[\|\boldsymbol{\mu}_k - \boldsymbol{e}_i\| - \delta_v]_+^2 \tag{1}$$

$$L_{push} = \frac{1}{C^2}\sum_{k=1}^{C}\sum_{l=1}^{C}[2\delta_d - \|\boldsymbol{\mu}_k - \boldsymbol{\mu}_l\|]_+^2 \tag{2}$$

where $\|\cdot\|$ is the L2-norm and $[x]_+ = max(0, x)$ is the rectifier function. The pull force L_{pull} (Eq. 1) brings the object's pixel embeddings closer to their mean embedding $\boldsymbol{\mu}_k$, while the push force L_{push} (Eq. 2) pushes the objects away, by increasing the distance between mean instance embeddings. Note that both terms are hinged, i.e. embeddings within the δ_v-neighbourhood of the mean embedding $\boldsymbol{\mu}_k$ are no longer pulled to it. Similarly, mean embeddings which are further apart than $2\delta_d$ are no longer repulsed.

We exploit the clustering induced by this loss to select pixels belonging to a single instance and apply auxiliary losses at the instance level (Fig. 2). Crucially, we find that given an instance I_k it is possible to extract a "soft" mask S_k for the current network prediction of the instance I_k in a *differentiable* way (Fig. 1). We select an anchor point for I_k at random and project it into the learned embedding space to recover its embedding $\boldsymbol{a}_k$, which we term anchor embedding. We compute the distance map from all image pixel embeddings to the anchor embedding and apply a Gaussian kernel function $\phi\colon \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}$ to "softly" select the pixels within the δ_v-neighborhood of $\boldsymbol{a}_k$ (δ_v is the pull term margin in Eq. 1):

$$\begin{aligned} S_k &= \{\phi(\boldsymbol{e}_i, \boldsymbol{a}_k) \mid i = 1, ..., N\} \\ \phi(\boldsymbol{e}_i, \boldsymbol{a}_k) &= \exp\left(-\frac{\|\boldsymbol{e}_i - \boldsymbol{a}_k\|^2}{2\sigma^2}\right) \end{aligned} \tag{3}$$

[1] Similarly to [6] a regularization term ($\frac{1}{C}\sum_{k=1}^{C}\|\mu_k\|$) which keeps the embeddings bounded is added to the final loss with a small weight of 0.001. For clarity, we omit this term in the text

We require the embeddings within the distance δ_v from the anchor embedding $\boldsymbol{a}_k$ have a kernel value greater than a predefined threshold $t \in (0, 1)$, i.e. $\phi(\boldsymbol{e}_i, \boldsymbol{a}_k) \geq t \iff \|\boldsymbol{e}_i - \boldsymbol{a}_k\| \leq \delta_v$. We can thus determine σ^2: substituting $\|\boldsymbol{e}_i - \boldsymbol{a}_k\| = \delta_v$ in Eq. 3, we get $\exp\left(-\frac{\delta_v^2}{2\sigma^2}\right) = t$, i.e. $\sigma^2 = \frac{-\delta_v^2}{2 \ln t}$. We choose $t = 0.9$ in our experiments and refer to Appendix A.7 for a detailed exploration of this hyperparameter.

We can now formulate a loss for a single object segmentation mask by minimizing the Dice loss (D) [36] of the mask S_k predicted using Eq. 3 and the corresponding groundtruth mask I_k.

$$L_{obj} = \frac{1}{C} \sum_{k=1}^{C} D(S_k, I_k) \tag{4}$$

Combining the losses in Eq. 1, Eq. 2 and Eq. 4, we get:

$$L_{SO} = \alpha L_{pull} + \beta L_{push} + \lambda L_{obj} \tag{5}$$

which we refer to as the Single Object contrastive loss (L_{SO}). We use $\alpha = \beta = 1$ (similar to [6]) and $\lambda = 1$ in our experiments. We set the pull and push margin parameters to $\delta_v = 0.5, \delta_d = 2.0$.

While Eq. 4 employs the Dice loss, our approach is not limited to Dice and can be used with any differentiable loss function at the single instance level, e.g. binary cross-entropy. Additionally, we explored the adversarial approach and trained a discriminator to distinguish the object masks coming from our differentiable instance selection method or from the groundtruth. More details can be found in Appendix A.1, the results are shown in Table 3.

3.2. Weak supervision

To enable training from positive unlabeled supervision, we introduce two additional loss terms: one to push each cluster away from the pixels in the unlabeled region and the other to enforce embedding space consistency in the unlabeled region. For an unlabeled region U which can contain both background and unlabeled instances, we define an additional "push" term:

$$L_{U\_push} = \frac{1}{C} \sum_{k=1}^{C} \frac{1}{N_U} \sum_{i=1}^{N_U} [\delta_d - \|\boldsymbol{\mu}_k - \boldsymbol{e}_i\|]_+^2 \tag{6}$$

where C is the number of labeled clusters/instances and N_U is the number of pixels in the unlabeled region U.

Since there is no direct supervision applied onto the unlabeled part of the image, the fully convolutional embedding network propagates the high frequency patterns present in there into the feature space. This is especially apparent for natural images and microscopy images with complex background structures, e.g. electron microscopy (see Fig 3 top left and Fig 6 top, col 3). To overcome this issue, we introduce the embedding consistency term. Given two different embedding networks f and g, we perturb the input image x with two different random, location- and shape-preserving augmentations t and t' and pass it through f and g respectively. The resulting vector fields $f(t(x))$ and $g(t'(x))$ come from the same input geometry, hence they should result in consistent instance segmentation after clustering, also in the unlabeled part of the input. To enforce this consistency we randomly sample an anchor point from the unlabeled region, project it into the $f-$ and g-embedding spaces, to get anchor embeddings $\boldsymbol{a}^f$ and $\boldsymbol{a}^g$ and compute two "soft" masks S^f and S^g according to Eq. 3. Similarly to Eq. 4 the embedding consistency is given by maximising the overlap of the two masks, using the Dice loss (D):

$$L_{U\_con} = \frac{1}{K} \sum_{k=1}^{K} D(S_k^f, S_k^g) \tag{7}$$

where K is the number of anchor points sampled from the unlabeled region U such that the whole region is covered by the union of extracted masks, i.e. $U \approx \bigcup_{k=1}^{K} S_k^f \cup S_k^g$. Having considered different variants of g-network including: weight sharing (with and without dropout) and independent training, we choose a momentum-based scheme [21, 20] where the network g (parameterized by θ_g) is implemented as an exponential moving average of the network f (parameterized by θ_f). The update rule for θ_g is given by: $\theta_g \leftarrow m\theta_g + (1-m)\theta_f$. f is trained by back-propagation. We refer to Appendix A.3 for extensive ablations of the g-network types and A.6 for the choice of a momentum coefficient $m \in [0, 1)$. Briefly, momentum variant provides the fastest convergence rate, improves training stability and is motivated by prior work [46, 8]. Significance of the embedding consistency term in weakly supervised setting is illustrated in Fig. 3. Note how the complex patterns present in the background (e.g. the flower pot) are propagated into the embedding space of the network trained without the consistency term (top, column 2), leading to spurious objects in the background after clustering (middle, column 2). In contrast, the same network trained with the embedding consistency loss results in crisp embeddings, homogeneous background embedding and clear background separation with no false positives (column 3). We confirm this observation by PCA-projecting the embeddings of background pixels onto 2D subspace (bottom). Network trained sparsely with the consistency term implicitly pulls background pixels into a single cluster, similar to the fully supervised network where the background pull is enforced by the loss. In contrast, the network trained without the consistency loss does not form a tight background cluster in the feature space. In addition, with a limited annotation budget of a certain number of objects, we achieve (see Appendix A.5) much better segmentation accuracy with objects distributed across many images

than with a few images fully labeled. The latter is prone to over-fitting, whereas a more diverse training set and the presence of the strong consistency regularizer in the former enables it to train from just a few object mask annotations. Our weakly supervised loss, termed Sparse Single Object loss (L_{SSO}), is given by:

$$L_{SSO} = \hat{L}_{SO} + \gamma \cdot L_{U\_push} + \delta \cdot L_{U\_con} \quad (8)$$

In our experiments we use $\gamma = \delta = 1$.

Tab 2 and Appendix A.2 shows that using the consistency term (Eq. 7) in a fully-supervised setting, in addition to the instance-based term (Eq. 4) improves the segmentation accuracy at the expense of longer training times. Fig. 2 gives a graphical overview of our training procedure which we term *SPOCO* (SParse Object COnsistency loss). Extensive ablation study of the individual loss terms can be found in Appendix A.2. In the experiments, we use the term SPOCO to refer to the fully-supervised training (taking all groundtruth objects including the background object). SPOCO@p refers to the weakly supervised positive unlabeled setting, in which a fraction $p \in (0, 1]$ of objects (excluding background) is taken for training. The background label is never selected in the weakly supervised setting, i.e. SPOCO@1.0 means that the network was trained with all labeled objects, excluding background.

3.3. Clustering

The final instance segmentation is obtained by clustering the predicted pixel embeddings. Mean-shift [13] and HDBSCAN [7] are commonly used for this task [6, 26, 41]. In this work we experimented with two additional clustering schemes: (1) partitioning [50] of a metric graph derived from pixel embeddings [28] and (2) a hybrid approach where initial mean-shift clusters are refined to conform with the pull-push loss formulation (Sec 3.1). Embeddings from networks f and g are used together in (2), all other clustering methods use the f-embeddings only. The advantage of (1) is a much faster inference time, whereas (2) can result in higher-quality segmentations. We refer to Appendix A.4 for a detailed description and comparison of different clustering methods.

4. Experiments

The fully- and semi-supervised evaluation is based on:

CVPPP. We use the A1 subset of the popular CVPPP dataset [37] which is part of the LSC competition. The task is to segment individual leaf instances of a plant growing in a pot. The dataset consists of 128 training images with public groundtruth labels and 33 test images with no publicly available labels. Test images come with a foreground mask which can be used during inference.

Cityscapes. We use Cityscapes [14] to demonstrate the performance of our method on a large-scale instance-level segmentation of urban street scenes. There are 2975 training images, 500 validation images, and 1525 test images with fine annotations. We choose 8 semantic classes: *person, rider, car, truck, bus, train, motorcycle, bicycle* and train the embedding networks separately for each class using the training set in the full and weak supervision setting. The object sampling procedure used for weakly supervised training is described in the supplementary.

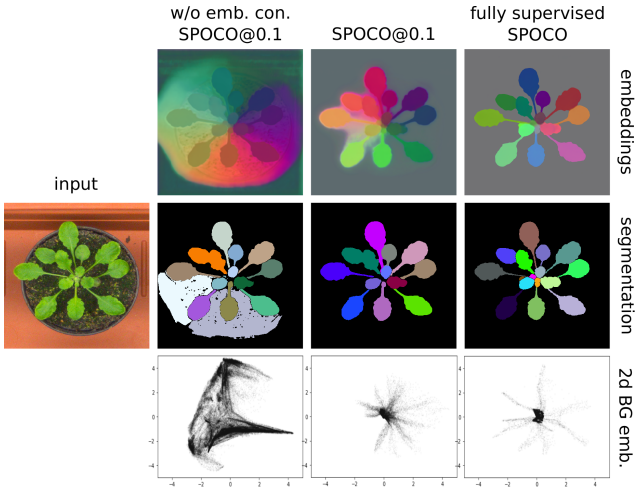

Figure 3. Different training schemes, left to right: SPOCO@0.1 trained without embedding consistency; SPOCO@0.1 trained with embedding consistency; SPOCO trained with full supervision (including the background label). **TOP)** PCA-projected embeddings; **MIDDLE)** corresponding clustering results; **BOTTOM)** background pixel embeddings PCA-projected onto 2D subspace.

Light microscopy (LM) datasets. To evaluate the performance of our approach on a challenging boundary-based segmentation task we selected a 3D LM dataset of the ovules of *Arabidopsis thaliana* from [51], with 48 image stacks in total: 39 for training, 2 for validation and 7 for testing. Additionally, we use the 3D *A. thaliana* apical stem cells from [49] in a transfer learning setting. The images are from the same imaging modality as the ovules dataset (confocal, cell membrane stained), but differ in cell type and image acquisition settings. We choose the Ovules dataset as the source domain and Stem cells as the target (*plant1, plant2, plant4, plant13, plant15* are used for fine-tuning and *plant18* for testing).

Electron microscopy (EM) datasets. Here we test our method in the transfer learning setting on the problem of mitochondria segmentation. An important difference between light and electron microscopy from the segmentation perspective lies in the appearance of the background which is simply dark and noisy for LM and highly structured for EM. The source domain (VNC dataset) [19] is a small annotated $20 \times 1024 \times 1024$px volume of the Drosophila larva VNC acquired with voxel size of $50 \times 5 \times 5$nm. We use 13 consecutive slices for training and keep 7 slices for validation.

As target domain we use the 3D MitoEM-R dataset from the MitoEM Challenge [48] a $500 \times 4096 \times 4096$px volume at $30 \times 8 \times 8$ nm resolution extracted from rat cortex. Slices (0-399) are used for fine-tuning and (400-499) for testing.

4.1. Setups

Any fully convolutional architecture with dense outputs could be used as an embedding network. We choose the U-Net [43, 12]. The depth of the U-Net is chosen such that the receptive field of features in the bottleneck layer is greater or equal to the input patch size. In all experiments we train the networks from scratch without using any pre-trained backbones. We use the Adam [25] optimizer with initial learning rate 0.0002 and weight decay 0.00001. Data augmentation consists of random crops, random flips, random scaling and random elastic deformations. For the momentum contrast embedding network, we additionally use additive Gaussian noise, Gaussian blur and color jitter as geometry preserving transformations.

In transfer learning experiments, the source network is always trained with the full groundtruth. On the target domain, we reduce the learning rate by a factor of 10 compared to the source network and use only a small fraction of the objects. VNC dataset is too small to train a 3D U-Net, so we perform EM segmentation in 2D, slice-by-slice. We also downsample VNC dataset by factor 1.6 in XY to match the voxel size of the target MitoEM data.

A detailed description of the network architecture, training procedure and hyperparameter selection can be found in the Appendix A.1.

4.2. Results and Discussion

CVPPP. Table 1 shows the results on the test set. The challenge provides foreground masks for test set images and we assume they have been used by authors of [6, 42, 26] in test time inference. In this setting, SPOCO outperforms [26] and the current winner of the leaderboard on the A1 dataset, keeping the advantage even in the case when the foreground mask is not given, but learned by another network ("predicted FG"). Even without using the foreground mask in the final clustering, SPOCO is close to [26] in segmentation accuracy, achieving much better average difference in counting score ($|DiC|$). We evaluate weakly supervised predictions without the foreground mask as we cannot easily train a semantic network without background labels. Nevertheless, even when training with only 10% of the groundtruth instances (SPOCO@0.1), the Symmetric Best Dice (SBD) as compared with the fully supervised SPOCO (without FG) drops only by 10 percent points. Qualitative results from SPOCO@0.1 can be seen in Fig. 3 (column 3), where the single under-segmentation error is present in the top left part of the image. HDBSCAN with $min\_size = 200$ is used for clustering in this case. Visual results and performance metrics for other clustering methods can be found in Appendix A.4. CVPPP dataset was used extensively in the ablation study, see A.2 for details.

Method	SBD	$\lvert DiC \rvert$
Discriminative loss [6]	0.842	1.0
Recurrent attention [42]	0.849	**0.8**
Harmonic Emb. [26]	0.899	3.0
SPOCO (GT FG)	**0.932**	1.7
SPOCO (pred FG)	0.920	1.6
SPOCO (w/o FG)	0.886	1.3
SPOCO@0.1	0.788 ± 0.017	5.4 ± 0.3
SPOCO@0.4	0.824 ± 0.003	3.2 ± 0.5
SPOCO@0.8	0.828 ± 0.010	1.6 ± 0.2

Table 1. Results on the CVPPP test set. Segmentation (SBD) and counting ($|DiC|$) scores for fully supervised SPOCO are reported in 3 different clustering settings: (1) with the groundtruth foreground mask, (2) with the predicted foreground mask (3) without the foreground mask. Results for semi-supervised setting SPOCO@p (no foreground mask) are presented for 10%, 40% and 80% of randomly selected groundtruth instances.

Cityscapes. We train our method with sparse (SPOCO@0.4) and full supervision and compare it with the fully-supervised contrastive framework [6]. In [6] authors trained a single model with multiple classes, applying the loss only within a given semantic mask. Since groundtruth semantic masks are not available when training from sparsely labeled instances, we train one model (including our implementation of [6]) for each semantic class. For inference we use pre-trained semantic segmentation model (DeepLabV3 [10]) to generate semantic masks and cluster the embeddings only within a given semantic mask. After initial mean-shift clustering we merge every pair of clusters if the mean cluster embeddings are closer than δ_d (push force hinge in Eq. 2). Average Precision at 0.5 intersection-over-union computed on the validation set can be found in Table 2. Our method outperforms [6] with only 40% of the groundtruth objects of each semantic class used for training. This is true for all classes apart from person, car and bicycle where the model requires larger number of annotated objects to reach high precision. Importantly, using consistency term in the fully-supervised setting improves the score by a large margin. The performance of SPOCO@0.4 is almost as good as the fully-supervised SPOCO without the consistency term. We hypothesize that strong regularization induced by the consistency term is crucial for classes with small number of instances. Fig. 4 shows qualitative results on a few samples from the validation set. Network trained with discriminative loss frequently over-segments large instances (trucks, buses, trains). A common mistake in crowded scenes for both methods is the merging of neighboring instances. Segmentation scores at different sampling rates, comparison with a class-agnostic training setting as

Class	DL [6]	S@0.4	S w/ con	S w/o con
person	**0.275**	0.230	0.260	0.270
rider	0.392	0.396	**0.451**	0.448
car	**0.416**	0.301	0.331	0.363
truck	0.486	0.558	**0.604**	0.527
bus	0.504	0.601	**0.637**	0.530
train	0.375	0.594	**0.656**	0.490
motorcycle	0.382	0.405	**0.464**	0.461
bicycle	**0.267**	0.214	0.266	0.255
average	0.387	0.412	**0.459**	0.418

Table 2. Segmentation results on the Cityscapes validation set. Average and per-class AP@0.5 scores are reported. *DL* - discriminative loss [6], *S@0.4* - SPOCO@0.4, *S w/ con* - fully-supervised SPOCO with the consistency term, *S w/o con* - fully-supervised SPOCO without the consistency term.

well as qualitative results can be found in the Appendix A.8.

Figure 4. Segmentation results for randomly selected images of different semantic classes on the Cityscapes validation set.

3D LM datasets. We compare SPOCO to the method of [51]: a 3-step pipeline of boundary prediction, supervoxel generation and graph agglomeration. Following [51], Adapted Rand Error [15] is used for evaluating the segmentation accuracy. As shown in Table 3, the performance of SPOCO is close to that of the much more complex 3-step PlantSeg pipeline. An additional adversarial loss term (SPOCO with L_{wgan}, see Appendix A.1) brings another performance boost and improves SPOCO accuracy beyond the [51] level.

Note that SPOCO trained with 10% of the groundtruth instances already outperforms the original embedding network with discriminative loss [6]. See Fig. 5 (top row) for qualitative results on a randomly sampled test set patch.

Table 4 shows SPOCO performance in a transfer learning setting, when fine-tuning a network trained on the Ovules dataset to segment the Stem dataset. The Ovules network trained only on source data does not perform very well, but just 5% of the target groundtruth annotations brings a two-fold improvement in segmentation accuracy. Results in tables 3 and 4 are based on HDBSCAN ($min\_size = 550$) clustering.

Method	ARand error
PlantSeg [51]	0.046
Discriminative loss [6]	0.074
SPOCO	0.048
SPOCO with L_{wgan}	**0.042**
SPOCO@0.1	0.069
SPOCO@0.4	0.060
SPOCO@0.8	0.057

Table 3. Evaluation on a 3D LM dataset of Arabidopsis Ovules [51]. The Adapted Rand Error (ARand error) averaged over the 7 test set 3D stacks is reported. Bottom part of the table shows the scores achieved in weakly supervised settings.

Method	ARand error
Stem only	0.074
Ovules only	0.227
Ovules+Stem@0.01	0.141 ± 0.002
Ovules+Stem@0.05	0.109 ± 0.002
Ovules+Stem@0.1	0.106 ± 0.004
Ovules+Stem@0.4	0.093 ± 0.003

Table 4. Evaluation on a 3D LM dataset in a transfer learning setting. Ovules dataset acts as the source domain, Stem dataset as target. Performance is shown as Adapted Rand Error, lower is better. Mean $\pm$ SD are reported across 3 random samplings of the instances from the target dataset.

Qualitative results are shown in Fig. 5 (bottom row). Note how the output embeddings from the Ovules network fine-tuned with just 1% of cells from the target dataset are less crisp due to the domain gap, but the clustering is still able to segment them correctly.

EM datasets. Table 5 continues the evaluation of SPOCO performance in a transfer learning setting. We report the average precision at 0.5 IoU threshold (AP@0.5) and the mean average precision (mAP). Similar to the LM case, just 1% of annotated objects in the target dataset bring a 1.5 fold improvement in the mean average precision compared to the network trained on source VNC domain only. A comparison to a network trained only on MitoEM (Table 5 bottom) shows that fine-tuning does significantly improve performance for low amounts of training data (1% of the target). With 10% of the annotated objects, fine-tuned VNC network does not reach the performance of the SPOCO@0.1 trained directly on MitoEM (target) due to reduced learning rate. Figure 6 illustrates the EM experiments. The VNC-net only partially recovers 4 out of 7 groundtruth instances and also produces a false positive. MitoEM@0.05 without consistency loss only recovers 2 instances, while the version with the consistency loss recovers

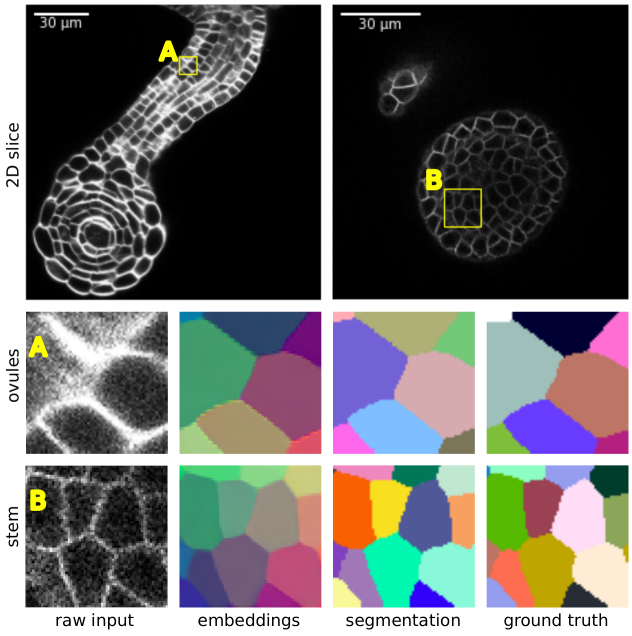

Figure 5. LM segmentation in standard and transfer learning settings. **TOP)** samples from the 3D Ovules (left) and Stem (right) datasets; **MIDDLE)** segmentation of a selected patch (A) from the source domain; **BOTTOM)** output of the source (Ovules) network fine-tuned with 1% of instances from the target (Stem) and the corresponding segmentation of a selected patch (B).

the correct segmentation. Embeddings clustered with HDBSCAN ($min\_size = 600$).

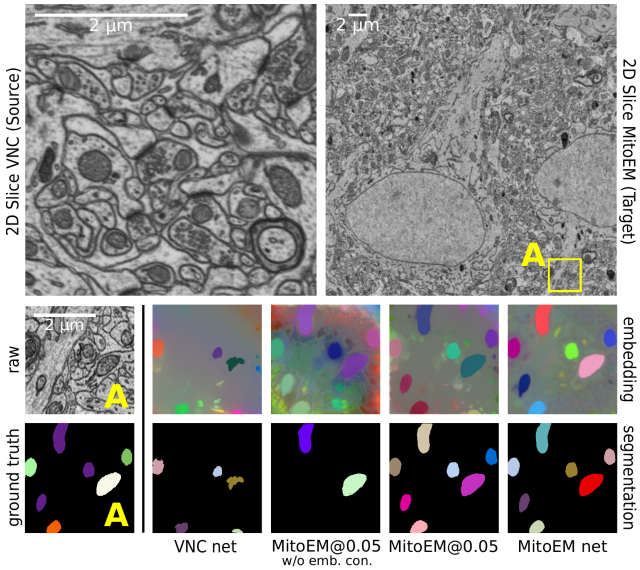

Figure 6. EM segmentation in transfer learning setting. **TOP)** samples from the source (VNC, left) and target (MitoEM, right) datasets; **MIDDLE)** the input image and the RGB-projected embeddings: trained on VNC only, VNC-pretrained + MitoEM@0.05-finetuned without the embedding consistency, same but with the embedding consistency, trained on MitoEM only; **BOTTOM)** groundtruth and predicted segmentations.

Method	AP@0.5	mAP
VNC	0.234	0.137
VNC-MitoEM		
SPOCO@0.01	0.368 ± 0.022	0.247 ± 0.022
SPOCO@0.05	0.398 ± 0.007	0.277 ± 0.006
SPOCO@0.10	0.389 ± 0.013	0.268 ± 0.007
MitoEM		
SPOCO@0.01	0.088 ± 0.045	0.046 ± 0.025
SPOCO@0.05	0.403 ± 0.055	0.280 ± 0.046
SPOCO@0.10	0.481 ± 0.008	0.340 ± 0.007
SPOCO	0.560	0.429

Table 5. Evaluation on MitoEM dataset (target) fine-tuned from the VNC net (upper part) and trained from scratch (lower part). The performance is measured through average precision (AP@0.5, mAP). Mean $\pm$ SD are reported across 3 random samplings of the instances from the target dataset.

5. Conclusion

We presented a novel approach to weak supervision for instance segmentation tasks which enables training in a positive unlabeled setting. Here, only a subset of object masks are annotated with no annotations in the background and the loss is applied directly to the annotated objects via a differentiable instance selection step. The unlabeled areas of the images contribute to the training through an instance-level consistency loss.

We demonstrate the advantage of single-instance losses in a fully supervised setting, reaching state-of-the-art performance on the CVPPP benchmark and improving on strong baselines in several microscopy datasets. Weak positive unlabeled supervision is evaluated on the Cityscapes instance segmentation task and on biological datasets from light and electron microscopy, 2D and 3D, in direct training and in transfer learning. In all cases, the network demonstrates strong segmentation performance at a very reduced manual annotation cost.

In the future, we plan to explore the possibility of fully self-supervised pre-training using the consistency loss and an extended set of augmentations. This would open up the possibility for efficient fine-tuning of the learned feature space with point supervision for both semantic and instance segmentation tasks.

Limitations. The main drawback of the proposed approach is the lack of a universal clustering method to assign instance labels to pixels based on their embeddings. The existing methods all have benefits and drawbacks; there is no consistent winner that would work robustly across all segmentation benchmarks. Appendix A.4 contains a detailed comparison of different clustering algorithms.

References

[1] T. Beier, C. Pape, N. Rahaman, T. Prange, S. Berg, D. D. Bock, A. Cardona, G. W. Knott, S. M. Plaza, L. K. Scheffer, et al. Multicut brings automated neurite segmentation closer to human performance. *Nature methods*, 14(2):101–102, 2017. 3

[2] J. Bekker and J. Davis. Learning from positive and unlabeled data: A survey. *Machine Learning*, 109(4):719–760, 2020. 3

[3] M. Bellver, A. Salvador, J. Torres, and X. Giro-i Nieto. Mask-guided sample selection for semi-supervised instance segmentation. *Multimedia Tools and Applications*, 79(35):25551–25569, 2020. 1

[4] M. Bellver Bueno, A. Salvador Aguilera, J. Torres Viñals, and X. Giró Nieto. Budget-aware semi-supervised semantic and instance segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019*, pages 93–102, 2019. 1

[5] T. Bepler, A. Morin, M. Rapp, J. Brasch, L. Shapiro, A. J. Noble, and B. Berger. Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs. *Nature methods*, 16(11):1153–1160, 2019. 3

[6] B. D. Brabandere, D. Neven, and L. V. Gool. Semantic instance segmentation with a discriminative loss function, 2017. 2, 3, 4, 5, 6, 7

[7] R. J. Campello, D. Moulavi, and J. Sander. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer, 2013. 5

[8] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. *CoRR*, abs/2104.14294, 2021. 4

[9] L. Chen, W. Zhang, Y. Wu, M. Strauch, and D. Merhof. Semi-supervised instance segmentation with a learned shape prior. In *Interpretable and Annotation-Efficient Learning for Medical Image Computing*, pages 94–102. Springer, 2020. 1

[10] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 6

[11] H. Cholakkal, G. Sun, F. Shahbaz Khan, and L. Shao. Object counting and instance segmentation with image-level supervision. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12389–12397, 2019. 1

[12] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. *CoRR*, abs/1606.06650, 2016. 6

[13] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 24(5):603–619, 2002. 5

[14] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2, 5

[15] CREMI. Cremi. miccai challenge on circuit reconstruction from electron microscopy images, 2017. https://cremi.org, 2017. 7

[16] A. Fathi, Z. Wojna, V. Rathod, P. Wang, H. O. Song, S. Guadarrama, and K. P. Murphy. Semantic instance segmentation via deep metric learning, 2017. 3

[17] J. Funke, F. D. Tschopp, W. Grisaitis, A. Sheridan, C. Singh, S. Saalfeld, and S. C. Turaga. A deep structured learning approach towards automating connectome reconstruction from 3d electron micrographs. *arXiv preprint arXiv:1709.02974*, 2017. 3

[18] W. V. Gansbeke, S. Vandenhende, S. Georgoulis, and L. V. Gool. Unsupervised semantic segmentation by contrasting object mask proposals, 2021. 3

[19] S. Gerhard, J. Funke, J. Martel, A. Cardona, and R. Fetter. Segmented anisotropic sstem dataset of neural tissue, 2013. 5

[20] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, B. Piot, k. kavukcuoglu, R. Munos, and M. Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc., 2020. 4

[21] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 4

[22] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2

[23] P. Hirsch, L. Mais, and D. Kainmueller. Patchperpix for instance segmentation. *arXiv preprint arXiv:2001.07626*, 2020. 3

[24] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 876–885, 2017. 2, 3

[25] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 6

[26] V. Kulikov and V. Lempitsky. Instance segmentation of biological images using harmonic embeddings, 2020. 5, 6

[27] I. H. Laradji, N. Rostamzadeh, P. O. Pinheiro, D. Vazquez, and M. Schmidt. Proposal-based instance segmentation with point supervision. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 2126–2130, 2020. 3

[28] K. Lee, R. Lu, K. Luther, and H. S. Seung. Learning and segmenting dense voxel embeddings for 3d neuron reconstruction. *IEEE Transactions on Medical Imaging*, pages 1–1, 2021. 5

[29] K. Lee, J. Zung, P. Li, V. Jain, and H. S. Seung. Superhuman accuracy on the snemi3d connectomics challenge. *arXiv preprint arXiv:1706.00120*, 2017. 3

[30] L. Lejeune and R. Sznitman. A positive/unlabeled approach for the segmentation of medical sequences using point-wise supervision, 2021. 1

[31] L. Lejeune and R. Sznitman. A positive/unlabeled approach for the segmentation of medical sequences using point-wise supervision. *Medical image analysis*, 73:102185, 2021. 3

[32] Q. Li, A. Arnab, and P. H. Torr. Weakly-and semi-supervised panoptic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 102–118, 2018. 2

[33] D. Lin, J. Dai, J. Jia, K. He, and J. Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3159–3167, 2016. 2, 3

[34] B. Liu, Y. Dai, X. Li, W. Lee, and P. Yu. Building text classifiers using positive and unlabeled examples. In *Third IEEE International Conference on Data Mining*, pages 179–186, 2003. 1

[35] B. Liu, Y. Dai, X. Li, W. S. Lee, and P. S. Yu. Building text classifiers using positive and unlabeled examples. In *Third IEEE International Conference on Data Mining*, pages 179–186. IEEE, 2003. 3

[36] F. Milletari, N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation, 2016. 4

[37] M. Minervini, A. Fischbach, H. Scharr, and S. A. Tsaftaris. Finely-grained annotated datasets for image-based plant phenotyping. *Pattern Recognition Letters*, pages –, 2015. 2, 5

[38] D. Neven, B. D. Brabandere, M. Proesmans, and L. V. Gool. Instance segmentation by jointly optimizing spatial embeddings and clustering bandwidth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[39] K. Nishimura, R. Bise, et al. Weakly supervised cell instance segmentation by propagating from detection response. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 649–657. Springer, 2019. 2

[40] C. Pape, A. Matskevych, A. Wolny, J. Hennies, G. Mizzon, M. Louveaux, J. Musser, A. Maizel, D. Arendt, and A. Kreshuk. Leveraging domain knowledge to improve microscopy image segmentation with lifted multicuts. *Frontiers in Computer Science*, 1:6, 2019. 3

[41] C. Payer, D. Štern, M. Feiner, H. Bischof, and M. Urschler. Segmenting and tracking cell instances with cosine embeddings and recurrent hourglass networks. *Medical Image Analysis*, 57:106–119, oct 2019. 5

[42] M. Ren and R. S. Zemel. End-to-end instance segmentation with recurrent attention, 2017. 6

[43] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3, 6

[44] K. Sofiiuk, O. Barinova, and A. Konushin. Adaptis: Adaptive instance selection network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7355–7363, 2019. 3

[45] C. Song, Y. Huang, W. Ouyang, and L. Wang. Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3136–3145, 2019. 2

[46] A. Tarvainen and H. Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*, 2017. 2, 4

[47] P. Vernaza and M. Chandraker. Learning random-walk label propagation for weakly-supervised semantic segmentation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017. 2, 3

[48] D. Wei, Z. Lin, D. Barranco, N. Wendt, X. Liu, W. Yin, X. Huang, A. Gupta, W. Jang, X. Wang, I. Arganda-Carreras, J. Lichtman, and H. Pfister. Mitoem dataset: Large-scale 3d mitochondria instance segmentation from em images. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, 2020. 6

[49] L. Willis, Y. Refahi, R. Wightman, B. Landrein, J. Teles, K. C. Huang, E. M. Meyerowitz, and H. Jönsson. Cell size and growth regulation in the arabidopsis thaliana apical stem cell niche. *Proceedings of the National Academy of Sciences*, 113(51):E8238–E8246, 2016. 5

[50] S. Wolf, C. Pape, A. Bailoni, N. Rahaman, A. Kreshuk, U. Köthe, and F. A. Hamprecht. The mutex watershed: Efficient, parameter-free image partitioning. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer Vision – ECCV 2018*, pages 571–587, Cham, 2018. Springer International Publishing. 5

[51] A. Wolny, L. Cerrone, A. Vijayan, R. Tofanelli, A. V. Barro, M. Louveaux, C. Wenzl, S. Strauss, D. Wilson-Sánchez, R. Lymbouridou, S. S. Steigleder, C. Pape, A. Bailoni, S. Duran-Nebreda, G. W. Bassel, J. U. Lohmann, M. Tsiantis, F. A. Hamprecht, K. Schneitz, A. Maizel, and A. Kreshuk. Accurate and versatile 3d segmentation of plant tissues at cellular resolution. *eLife*, 9:e57613, jul 2020. 5, 7