Decoupled Semantic Prototypes enable learning from diverse annotation types for semi-weakly segmentation in expert-driven domains

Simon Reiß[1] Constantin Seibold[1] Alexander Freytag[2] Erik Rodner[3] Rainer Stiefelhagen[1]
[1]Karlsruhe Institute of Technology [2]Carl Zeiss AG [3]University of Applied Sciences Berlin
{simon.reiss, constantin.seibold, rainer.stiefelhagen}@kit.edu,
alexander.freytag@zeiss.com, erik.rodner@htw-berlin.de

Abstract

A vast amount of images and pixel-wise annotations allowed our community to build scalable segmentation solutions for natural domains. However, the transfer to expert-driven domains like microscopy applications or medical healthcare remains difficult as domain experts are a critical factor due to their limited availability for providing pixel-wise annotations. To enable affordable segmentation solutions for such domains, we need training strategies which can simultaneously handle diverse annotation types and are not bound to costly pixel-wise annotations. In this work, we analyze existing training algorithms towards their flexibility for different annotation types and scalability to small annotation regimes. We conduct an extensive evaluation in the challenging domain of organelle segmentation and find that existing semi- and semi-weakly supervised training algorithms are not able to fully exploit diverse annotation types. Driven by our findings, we introduce Decoupled Semantic Prototypes (DSP) *as a training method for semantic segmentation which enables learning from annotation types as diverse as image-level-, point-, bounding box-, and pixel-wise annotations and which leads to remarkable accuracy gains over existing solutions for semi-weakly segmentation.*

1. Introduction

Modern semantic segmentation pipelines like [11,24,64] enable a wide range of segmentation applications in domains like urban scenes [13,41] or natural images [19,37,73]. Besides recent progress in supervised training and network architectures, an (even more?) important reason for this success was the availability of large budgets for the creation of training datasets. A significant portion of these budgets is usually spent on human labor for creating pixel-wise annotation masks. Motivated by cost reduction, crowd-sourced annotation with briefly instructed

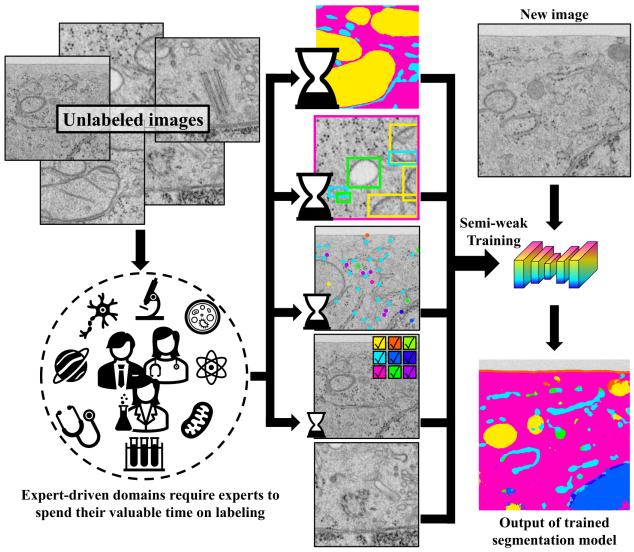

Figure 1. The availability of domain experts for annotating data is often the major bottleneck in expert-driven domains. Can we give experts the freedom to provide diverse annotation types based on their available time, and still train models successfully?

annotators became a popular choice, e.g., for the creation of MSCoco [37] or OpenImages [7]. However, this is often not possible for application domains where extensively trained experts need to provide annotations. As an example, Guay *et al.* highlighted in [23] that *"Cell biologists have used [...] segmentations of cellular structures to provide rich [...] new insights into cellular processes"*, but annotation *"required nine months' work from two in-lab annotators and represented a small fraction of all imaged cells."*.

For such expert-driven domains (Fig. 1), the time of skilled annotators needs to be used more efficiently, as their number can not be easily scaled up. Hence, training algorithms should not rely on naively extending collections of pixel-wise masks. Instead, training algorithms should be flexible towards the annotation-granularity which an expert has time to and is willing to provide (*expert-centrism*) and should still efficiently exploit small quantities of pixel-

wise annotations (*annotation efficiency*). Thereby, experts become the center of the segmentation pipeline, and can leverage their time and knowledge flexibly, *e.g*. in form of bounding box-, image-level- or single point annotations.

In this work, we evaluate both properties for existing training algorithms. As a starting point, we follow common semi-supervised scenarios and analyze *annotation efficiency* where we systematically reduce available pixel-wise annotations. In addition, we analyze existing training algorithms towards their *expert-centrism* by varying the number of pixel-wise masks and the annotation types. Based on these findings, we finally present *Decoupled Semantic Prototypes*: an expert-centric training algorithm applicable to any off-the-shelf segmentation architecture, designed by adapting ideas from contrastive learning to semi-weakly supervised scenarios and combining it with end-to-end trainable protype-based segmentation networks. Our experiments give insight into which algorithms can be flexibly used in expert-driven domains, and also give intuitions on how to spend annotation budgets: always on pixel-wise masks or on coarser annotations?

In summary, our contributions are as follows:

(1) We carefully benchmark segmentation training algorithms by investigating how well they scale to fewer and fewer pixel-wise annotations and how well they handle diverse weak label types, giving strong baselines and insights into their applicability for expert-driven domains.

(2) We introduce the *Annotation Compression Ratio* (ACR) as a metric for analyzing semi-weakly supervised algorithms towards their *annotation efficiency* and *expert-centrism* w.r.t. full supervision signals.

(3) We propose *Decoupled Semantic Prototypes* as a simple but efficient method for expert-centric semi-weakly segmentation which can learn from a diverse mix of annotation-types and which leads to gains of up to $+12.8\%$ absolute DICE for organelle segmentation.

2. Related Work

Semi-supervised Segmentation Several semi-supervised segmentation algorithms exploit a network trained on annotated data to predict pseudo-labels [30, 76] for unlabeled images during training to train with both image sets. One prominent example is FixMatch by Sohn *et al*. [51], which initially was introduced for classification and was transferred to segmentation in [50]. FixMatch produces pseudo-labels from mildly augmented images as supervision signal for strongly augmentation versions of the same image and often leads to strong results. Zhong *et al*. [71] use this idea of weak and strong augmentation in a contrastive learning (CL) framework. Combining self-supervision and CL on unlabeled data is done increasingly in semi-supervised learning [2, 65, 69, 75]. We will extend these ideas and adapt a specific CL variant, the simple yet significant idea of Decoupled Contrastive Learning (DCL) [66], to be applicable to diverse annotation types which traditional semi-supervised methods are not capable of.

Weakly-supervised Segmentation Pixel-wise masks are extremely costly to generate. As an example, Cordts *et al*. [13] report 1.5 hours for each pixel-wise annotation including quality control. Lin *et al*. [37] report about 15 minutes per pixel-annotatated image. To circumvent these massive efforts, a lot of work has flown into training with cheaper annotation-types. Most effort in weakly-supervised segmentation has been spent in the natural image domain using image-level labels [1, 26, 32, 43, 63, 67], bounding boxes [14, 28, 29, 34, 52] and scribbles, click- or point annotations [5, 6, 10, 35, 36, 56, 57, 60]. As an ill-posed problem, the task has been mostly addressed by finding useful assumptions and valid priors which can be exploited in training. Unfortunately, such assumptions are often designed for object-centric images, and hardly transfer to specialized, expert-driven application domains. Prominent techniques [20, 48, 72] may fail for domains substantially different from ImageNet [15]. Specialized weakly-supervised methods have been proposed for expert-driven domains [31, 49, 68]. In a similar spirit, we aim to exploit diverse weak annotation types and at the same time benefit from unlabeled images .

Semi-weakly Supervised Segmentation The previous two ideas are combined with semi-weak supervision [12] (also called mixed- or omni-supervision [21, 46, 62]). The idea is simple: exploit any available data during training, irrespective of being unlabeled, weakly labeled, or pixel-wise annotated. This makes training flexible towards different annotation types, thereby liberating application experts from being forced to provide time-consuming pixel-wise masks. The focus of semi-weakly supervision was mainly on pairs of annotation types, *e.g*. pixel-wise masks combined with image-level labels [4, 39, 40, 44], with bounding boxes [18, 27, 28, 36, 42, 55, 70], or with partial labels [17, 45]. More exotic combinations include unpaired masks [59] or unlabeled images [21] with scribbles. Li *et al*. [34] bootstrap panoptic segmentation from masks, image-level labels, and bounding boxes coupled with assumptions on natural images.

Inspired by these works, we will go one step further and enable semi-weakly supervised segmentation training with pixel-wise masks, bounding boxes, point annotations, image-level labels, and unlabeled images.

3. Decoupled Semantic Prototypes

Inspired by the previously reviewed methods, we now introduce *Decoupled Semantic Prototypes* (DSP). The design of DSP will allow us to exploit diverse annotation types,

hence, being expert-centric, while remaining applicable to off-the-shelf segmentation architectures.

3.1. Preliminaries

For training segmentation models, we define a training dataset as $\mathcal{D} = \{x_1, \dots, x_n | x_\ell \in \mathbb{R}^{c_{dim} \times H \times W}\}$ where images x_ℓ have c_{dim} color or intensity channels and H and W denote height and width. Since we do not want to restrict the training to particular annotation types, we allow for a broad semi-weakly supervised segmentation scenario where images x_ℓ may come unlabeled $\mathcal{U}$, with pixel-wise masks $\mathcal{M}$, or with weak annotations via bounding boxes $\mathcal{B}$, point annotations $\mathcal{P}$, or image-level labels $\mathcal{I}$. For an image x_ℓ that is labeled with a mask, we also have access to the weaker annotation types, as they can be derived from it. Similarly, box- and point-annotated images include image-level information as it can be directly derived.

3.2. Pixel-wise Embeddings

A key idea of our method is to integrate different annotation modalities by carefully designing dependencies on a pixel level. Therefore, each pixel needs to be associated with an embedding vector. In contrast to standard segmentation networks that are trained with a linear layer and a pixel-wise cross-entropy loss on top, we design our network ε such that $\varepsilon(x)$ yields an embedding $F \in \mathbb{R}^{D \times H \cdot W}$ consisting of embedding vectors $f_i \in \mathbb{R}^D$ for each pixel i. This adaptation can be easily done for all segmentation networks. For details about the network design used in our experiments see Section 5.2.

3.3. Semantic Prototype Association

Integrating and combining different annotation types requires modelling dependencies between the input data and class labels. Hence, in addition to the embeddings for each pixel, we further need an association to semantic classes. Therefore, we introduce semantic prototype vectors $p_c^j \in \mathbb{R}^D$, with c indicating the class represented by the prototype and j indexing the explicit prototype vector among a set P_c of prototypes for class c. To evaluate if a pixel-embedding f belongs to a class c, we first compute its similarity to every prototype p_c^j via the cosine distance:

$$\sigma(f, p_c^j) = \frac{f^\top p_c^j}{||f|| \cdot ||p_c^j||} \tag{1}$$

and then average the similarities between all prototypes of a given class c and the embedding f:

$$s_c(f, P_c) = \frac{1}{|P_c|} \cdot \sum_{j \in P_c} \sigma(f, p_c^j) \ , \tag{2}$$

thereby aggregating them to a class-wise score s_c. In training, these class-wise scores $s_c(f, P_c)$ are normalized by temperature scaling with τ and the softmax function:

$$\bar{s_c}(f, P_c) = \frac{\exp\left(s_c(f, P_c)\right)/\tau}{\sum_{i=1}^{C} \exp\left(s_c(f, P_i)/\tau\right)} \ . \tag{3}$$

The scores $\bar{s_c}$ can be plugged into the commonly used cross-entropy loss to train the segmentation network. At inference time, the class with highest class score $\arg\max_c \{s_c(f, P_c)\}_{c=1}^C$ gives the hard decision for each embedding vector $f \in F$, thereby giving the final semantic segmentation.

All prototypes p_c^j are parameters of the network and learned end-to-end (similar to [3]), or in pytorch-like code:

```
P = torch.randn((C, |P_C|, D))
P = P.div(P.pow(2).sum(2,keepdim=True).pow(0.5))
P = torch.nn.Parameter(P, requires_grad=True)
```

This can be thought of as implicitly finding cluster centers, which differentiates our prototypical segmentation from Zhou *et al*. [74], who require online clustering to acquire prototypes. Furthermore, class queries in transformers [54] are also coarsely related to this view on segmentation. To recap, rather than predicting one-hot categorical vectors, we learn data-driven D-dimensional class representations compatible to pixel embeddings. Using multiple prototypes per class allows for intra-class variability and multi-modal distributions in the learned embedding space (see supplement).

3.4. Decoupled Contrastive Prototypes

Modeling segmentation as prototype association provides us with the freedom to decide which associations between pixel-embeddings and prototype vectors to enforce. One way to enforce specific associations between sets of vectors is contrastive learning or in our case decoupled contrastive learning (DCL, [66]). Contrastive learning encourages that for a given image and its augmented version, the resulting representations z_i and $\hat{z}_i$ become similar by minimizing:

$$-\log \frac{\exp\left(\sigma(z_i, \hat{z}_i)/\tau\right)}{Z_i} \ , \tag{4}$$

with normalization Z_i computed over a batch of size B:

$$Z_i = \sum_{j=1, j \neq i}^{B} \exp\left(\sigma(z_i, \hat{z}_j)/\tau\right) + \exp\left(\sigma(z_i, z_j)/\tau\right) . \tag{5}$$

Different from standard contrastive learning, decoupled contrastive learning (DCL) removes the positive association of the numerator out of the denominator. We adapt DCL to not relate augmented vectors to each other, but to associate pixel-embeddings to semantic prototypes:

$$L(f_i, c) = -\log \frac{\exp\left(s_c(f_i, P_c)/\tau\right)}{Z_{i,c}} \tag{6}$$

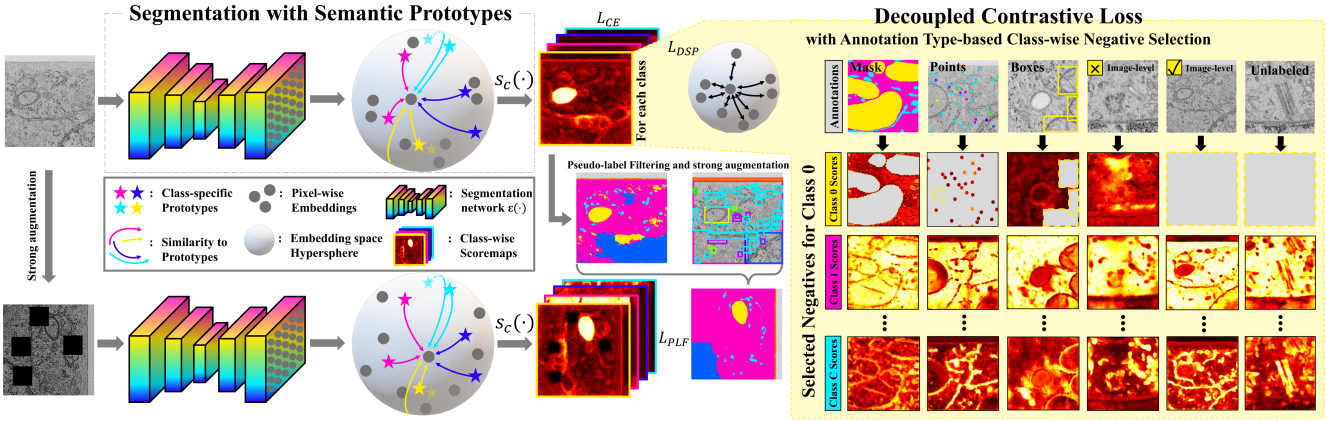

Figure 2. Left: Our segmentation method *Decoupled Semantic Prototypes* embeds each pixel of an input image into an embedding space, in which it assigns class-attribution by relating pixel-embeddings to learned class-specific prototypes. We add the idea of filtering self-inferred pseudo-labels with weak labels if available. Right: We extend the idea of Decoupled Contrastive training to learn from diverse annotation types by using them to decide which class-associations can be regarded as negatives and which should not (gray regions).

with $Z_{i,c}$ normalizing with respect to all pixels:

$$Z_{i,c} = \sum_{j=1}^{B\cdot H\cdot W} \sum_{k=1, k\neq c \wedge j\neq i}^{C} \exp\left(s_k(f_j, P_k)/\tau\right) \quad . \tag{7}$$

As we minimize Eq. (6), it encourages the association of f_i to prototypes in P_c to be high, and all other associations f_j to prototypes in P_c to be small. This is not desirable, as we would encourage the similarity between an arbitrary f_j and the elements in P_c to be small, even though pixel j could potentially belong to class c. Thus, we modify $Z_{i,c}$ as follows yielding our final formulation:

$$Z_{i,c} = \sum_{j=1}^{B\cdot H\cdot W} \sum_{k=1, c\notin\mathcal{A}_j \vee k\neq c}^{C} \exp\left(s_k(f_j, P_k)/\tau\right). \tag{8}$$

Here, $\mathcal{A}_j$ is a function returning the set of possible classes the pixel j might be associated with under the given annotation. For example, if j is a pixel drawn from an unlabeled image, then $\mathcal{A}(j)$ is the set of all classes, as we do not have any knowledge about classes for j. If the pixel j belongs to an image with image-level labels, $\mathcal{A}$ returns the set of these image-level classes. Similarly, $\mathcal{A}$ returns all classes of the bounding boxes in which pixel j is located in. For point annotations and pixel-wise masks, only a single label is relevant and $|\mathcal{A}(j)| = 1$.

By only using embedding-prototype associations $s_k(f_j, P_k)$ in the denominator (as negatives) that satisfy $c \notin \mathcal{A}(j) \vee k \neq c$, we make sure that pixel-embeddings and prototype pairs that are pushed apart through the contrastive term are guaranteed to encode different semantics. By utilizing decoupling, all embeddings that are associated to a class c share the same negatives, such that they only have to be computed once for each class in a batch, and not for each pixel-embedding, which would be computationally challenging. Note that our integration of semantics into embeddings and prototypes is flexible to any type of precise, partial, or ambiguous annotation that gives semantic cues whether a pixel belongs to a class.

3.5. Choosing Positive Associations

One notion missing in Equation (8) is the mechanism how positives (numerators) are determined. This is especially important, as this choice can introduce a considerable amount of faulty associations. As an example, imagine a thin, diagonally oriented object. In this case, most pixels in its bounding box do not belong to the class associated with the box. Simply taking all pixel-embeddings within the bounding box as positives for the box-class would introduce a lot of noise. Thus, we now discuss how positives can be carefully chosen for different annotation types.

Pixel-wise masks In case we have pixel-wise masks, we know the association between the pixel and its class and therefore between pixel-embedding and class-specific prototypes precisely. In this case, we propose to average pool all embedding-prototype associations within an instance (*i.e.*, a connected component in the mask) to represent the association of the whole instance to a class. This pooled embedding-prototype association serves as positive. The set of all pooled associations based on mask annotations for class c in the batch are denoted as Ω_c^m.

Point annotations Point annotations are also precise annotations, indicating the exact class a pixel-embedding should be associated to. As such, we simply take the pixel-embedding as it is and declare it as positive. The full set of point annotations of class c in the batch is Ω_c^p.

Image-level labels Positives based on an image-level label can be formed by means of a pooling function, as often used in multiple-instance learning. We use average pooling, *i.e.* for an image-level label containing class c,

we average pool the associations to c over the spatial dimensions of the whole embedding map F of the image: $\frac{1}{H \cdot W} \sum_{f_i \in F} s_c(f_i, P_c)$. We refer to Ω_c^{im} as the set of image-level pooled embedding associations to class c in the current batch.

Bounding boxes An implicit property of bounding boxes is that in each vertical- and each horizontal line of pixels in the box, at least one pixel corresponds to the class c of the box-label [58]. We use this property to pool the embedding-prototype associations for the box label c. As such, we take the maximum values along all vertical and horizontal directions and sum up all $s_c(f_i, P_c)$. Here, we assume that the maximum values are likely to lie on the object that is associated to the box class. We denote the set of bounding box-based positives as Ω_c^b.

Loss formulation Bringing everything together, our annotation type unified loss function can be written as:

$$L_{DSP} = \sum_{l \in \{m,b,p,im\}} \lambda_l \sum_{c=1}^{C} \sum_{f_i \in \Omega_c^l} L(f_i, c) \ , \qquad (9)$$

with weighting factors λ_l. Note that computing the loss can be simplified, since the normalization $\log(Z_{i,c})$ appears $|\Omega_c^l|$ times in Equation (9) (for details see supplement).

3.6. Decoupled Semantic Prototypes for Annotation Efficient and Expert-centric Segmentation

Given the prototype-based segmentation and decoupled contrastive loss function L_{DSP}, we can now outline the general training strategy for our *Decoupled Semantic Prototypes* method. First of all, we exploit a standard supervised cross-entropy loss L_{CE} using our temperature scaled softmax predictions Eq. (3) and the available masks.

Additionally, we integrate the commonly used idea of pseudo-labels [30, 51]: use a weakly augmented version of an image to compute pseudo-labels (PL) which are used as target for a strongly augmented version of the same input image. This way, we can learn from unlabeled data and regularize the model by training towards augmentation invariance. However, to make best use of weak labels, we integrate the idea of pseudo-label filtering for segmentation. Hence, instead of directly using predicted PL for weakly annotated images, we filter the pre-softmax scores as follows:
(i) **Image-level labels**: Set all class-predictions for class c to $-\infty$ if c is not in the image-level label.
(ii) **Bounding boxes**: Set all class-predictions for class c to $-\infty$ at pixels where no box of class c is located + (i).
(iii) **Point annotations**: Set all class-predictions at the point location to the class given by the point annotation + (i).

After pseudo-label filtering, we take the $\arg\max$ over the class dimension to get hard pixel-wise class assignments. We refer to the pseudo-label-based cross-entropy loss between the predictions on the strongly augmented input and the pseudo-label filtered target as L_{PLF}. We leverage pseudo-label filtering also to create strong baselines from semi-supervised literature [30,45,51] and extend them to the semi-weakly supervised scenario.

The total loss of our *Decoupled Semantic Prototypes* for semi-weakly and expert-centric segmentation is:

$$L_{total} = L_{CE} + L_{PLF} + L_{DSP}. \qquad (10)$$

The full proposed method is displayed in Figure 2 with our Decoupled Contrastive Loss on the right.

4. Standardized semi-weakly segmentation

Benefits of training approaches that tackle a lack of annotations is commonly investigated by training on increasing portions of fully labeled data [8, 33]. In scenarios with mixed annotation types, this is a rather simplistic form of *annotation-efficiency* and a new measure is required.

Annotation Compression Ratio (ACR) Given a dataset which is completely labeled with a *base annotation type*, a training regime leveraging a small portion of these base annotations can be viewed as compressing them. Inspired by this perspective, we define the *Annotation Compression Ratio (ACR)* of an algorithm on a dataset with respect to a base annotation type (*e.g.* pixel-wise annotations) as

$$\text{ACR} = \frac{\text{\# total base annotations}}{\text{\# used base annotations}} \ . \qquad (11)$$

Thus, an algorithm that has been trained with an ACR $= 2$ cuts the used base annotations in half (*cf.* with [9] for a similar notation used for neural network pruning). Since accuracy as a function of the number of annotated images is generally regarded to follow power laws [53], we propose to exponentially sample ACR values, *i.e.*, subsequently cutting the number of base annotations in half.

For reducing pixel-wise masks $\mathcal{M}$ (base annotation), we substitute it with weak labels ($\mathcal{B}$, $\mathcal{P}$, $\mathcal{I}$) or no labels ($\mathcal{U}$) and leverage them in training – yielding what we refer to as a semi-weakly supervised training scenario.

5. Evaluation

5.1. Experimental Setup

Dataset We evaluate semi-weakly supervised segmentation algorithms on the challenging OPENORGANELLE data collection by Heinrich *et al.* [25]. We focus on the four cell datasets *HELA-2*, *HELA-3*, *JURKAT-1*, *MACROPHAGE-2* due to their difficulty and diversity. These datasets are large tissue volumes scanned with focused ion beam scanning electron microscopes (FIB-SEM) and come with annotated sub-volumes. The segmentation task is to segment cell organelles in these sub-volumes, which are processed as 2D slices. For a statistically sound analysis, we create cross-validation splits via cross-sub-volume train/validation/test

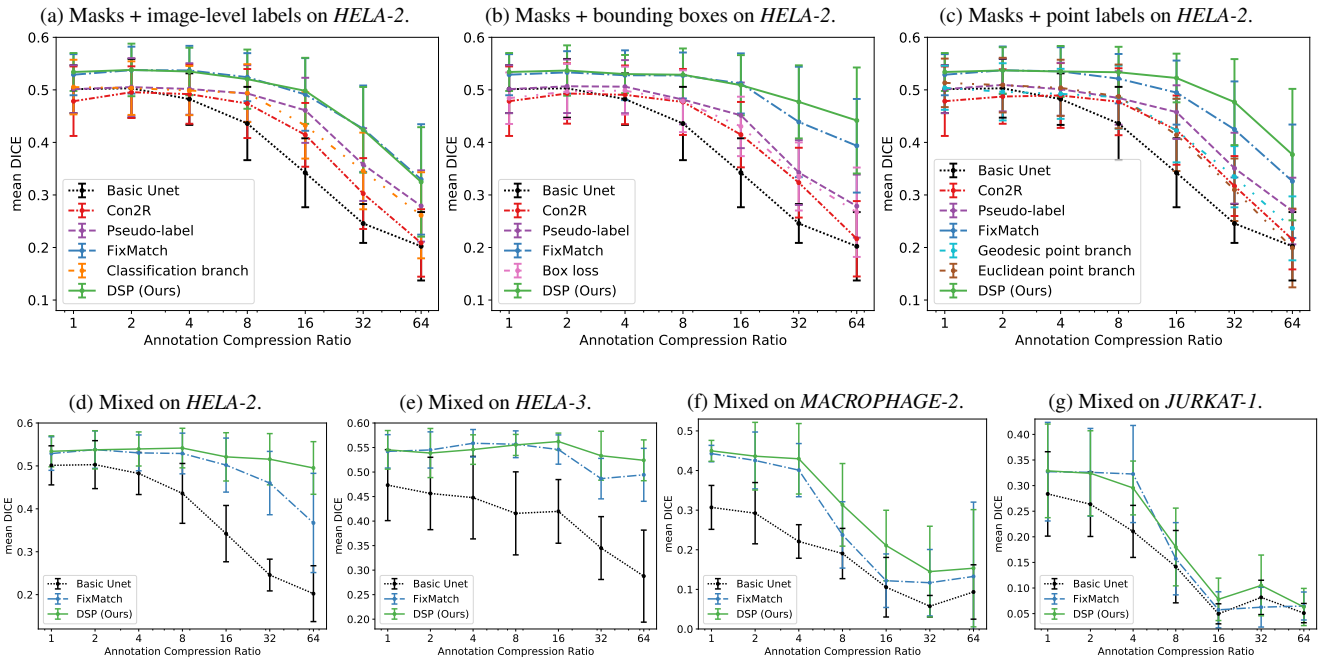

Figure 3. Segmentation accuracy as a function of reduced annotations, measured as mean and standard deviation in mean DICE for seven ACRs. Note different scales on y-axes. Top row: Semi-weakly supervised segmentation algorithms for *pairs* of annotation types on the *HELA-2* data. Results are obtained from 10 cross-validation splits. Bottom row: Results on *HELA-2*, *HELA-3*, *MACROPHAGE-2* and *JURKAT-1* trained with five annotation types mixed. Results obtained from 10, 5, 5, 5 splits respectively. Numerical results in supplement.

splits under the side-condition that every class is present in at least one sub-volume per split. However, since many of the OPENORGANELLE classes are highly specialized, this condition is rarely fulfilled. Therefore, we merge classes into 17 classes following a biologically consistent class-hierarchy (*e.g.*, merging *mitochondria*, *mitochondria membrane* and *mitochondria DNA*). Rare classes occurring in less than three sub-volumes are excluded due to the requirement for cross-sub-volume validation. This results in 11 classes for *HELA-2*, 10 for *HELA-3*, 8 for *JURKAT-1*, and *MACROPAHGE-2*. In total, we obtain 10 cross-validation splits for the largest dataset *HELA-2* and 5 for the remaining ones. Each split is randomly shuffled, with the exception that all C classes need to be present in the first C images. Finally, we ensure that the annotated images for small ACRs contain all annotations of large ACRs.

Semi-weakly Supervision Scenarios In all our experiments, we reduce the amount of costly pixel-wise annotations exponentially, *i.e.*, we double the ACR successively from 1 to 64 (which reduces the fraction of images with pixel-wise masks from 100% to 1.6%). We first evaluate three semi-weak scenarios where these pixel-wise annotations are combined with image-level labels, with point labels, or with bounding boxes. As mentioned in Sec. 4, we analyze scenarios where these supervision-types are available for all images without pixel-wise masks. Finally, the most important scenario for expert-centric and annotation efficient segmentation is the availability of diverse annotation types. We simulate this scenario by distributing all four coarse annotation types $\mathcal{B}$, $\mathcal{P}$, $\mathcal{I}$, $\mathcal{U}$ uniformly among images without pixel-wise masks (*e.g.*, for $ACR = 2$: 50% pixel-wise masks, 12.5% unlabeled, 12.5% image-level labels, 12.5% point annotations, 12.5% bounding boxes). We always report mean and standard deviation of average class-wise DICE scores over the cross-validation splits for all four datasets and for seven exponentially distributed ACRs.

Implementation details In our experiments, all methods are implemented with the same Unet architecture [47]. Although all methods are applicable to other segmentation architectures as well, we intentionally chose Unets due to their stability such that side-effects like missing learning rate warmup are not to be expected. All models are trained with AdamW [38] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, a learning rate of $6e^{-5}$, weight decay equal to 0.01, and Xavier initialization [22]. All trainings are performed on a multi-GPU setup with 4× 40GB NVIDIA A100-40 for 100 epochs on each split. For each split, validation is done every 10 epochs, and each val-best model is evaluated on the corresponding test set after training. As different training methods have different memory requirements, we always set the batch size to the maximally possible size under the method's memory consumption (between 16 and 28). Batching requires equally-sized inputs, but the datasets have varying image sizes. Thus, we zero-pad all images to the respective maximal size. For weak augmentation, we use horizontal and vertical flipping, rotations by $0°, 90°, 180°, 270°$, jitter

brightness, contrast, saturation and hue by a factor of 0.2. Strong augmentations add flipping and rotating, and applying a jitter factor of 0.4 and CutOut [16] up to nine times. Code available at https://github.com/Simael/DSP.

5.2. Baselines

We compare *DSP* with a variety of competing methods.

Basic Unet [47] being trained with L_{CE} exclusively from pixel-wise annotations is the naive baseline. This serves as initialization for all remaining methods.

Pseudo-label [30] is a standard semi-supervised segmentation baseline, where the network predicts labels for unlabeled images online and uses them for self-supervision. It is applicable as L_{PLF} to all semi-weakly supervised scenarios by adapting it with pseudo-label filtering (see Sec. 3.6).

Con2R [45] is a semi-weakly supervised method originally developed for 3D segmentation. We adapt Con2R as semi-weakly supervised method to 2D and set the receptive consistency size to 16×16. We notice an accuracy increase when sampling query- and neighbor sets from strongly- and weakly augmented views instead of just the strongly ones as in [45]. We add pseudo-label filtering directly to the semantic consistency constraint for semi-weakly segmentation.

FixMatch [51] was initially designed for classification and adapted for segmentation. Paired with pseudo-label filtering, FixMatch is a strong and versatile baseline for semi-weakly learning.

Classification branch as in [40] is used for the scenario of image-level- and pixel-wise labeled images. A dual-head architecture with segmentation output-head and classification branch serves as an obvious baseline.

Euclidean/Geodesic point branch is similar to [40] but for pixel-wise and point-labeled images. Since we are not aware of existing techniques for this combination, we obtain baselines by inspirations from interactive segmentation [61]. As auxiliary self-supervised task, we specify the second output head to regress point-based distance maps either for euclidean or geodesic distances (see supplement).

Box loss: For the scenario with pixel-wise annotations and boxes, we integrate the bounding box-based loss of Tian *et al*. [58] as alternative methods often rely on natural image priors which do not hold for many expert-centric domains.

DSP (Ours) In order to compute pixel-embeddings, we replace the final classification layer of a Unet with a sequence of batch norm, 1×1 convolutions with 64 kernels, LeakyReLU, and final 1×1 convolutions with 64 kernels. This plug-in-replacement produces embeddings with $D = 64$. We use 5 learned prototypes per class, temperature $\tau = 0.05$, and set the annotation type weights to $\lambda_m, \lambda_b, \lambda_p, \lambda_{im} = 0.1$.

L_{CE}	L_{PLF}	L_{DSP}				Architecture		DICE
		λ_m	λ_b	λ_p	λ_{im}	τ	$\lvert P_c \rvert$	
✓	✓	–	–	–	–	0.05	10	55.9%
✓	✓	1.0	1.0	1.0	1.0	0.05	10	58.4%
✓	✓	1.0	0.2	0.5	0.3	0.05	10	57.7%
✓	✓	0.1	0.1	0.1	0.1	0.05	10	59.5%
✓	✓	0.01	0.01	0.01	0.01	0.05	10	56.8%
✓	✓	0.1	0.1	0.1	0.1	1.0	10	41.9%
✓	✓	0.1	0.1	0.1	0.1	0.5	10	49.3%
✓	✓	0.1	0.1	0.1	0.1	0.01	10	58.6%
✓	✓	0.1	0.1	0.1	0.1	0.005	10	59.4%
✓	✓	0.1	0.1	0.1	0.1	0.05	1	59.0%
✓	✓	0.1	0.1	0.1	0.1	0.05	5	**60.7%**
✓	–	–	–	–	–	0.05	5	48.9%

Table 1. Ablation studies for *Decoupled Semantic Prototypes* on the first split of *HELA-2* mixed supervision at $ACR = 8$.

5.3. Quantitative results

Supervision $\mathcal{M} + \mathcal{I}$: For the scenario of image-masks and image-level labels, we observe in Fig. 3a a strong decline in mean DICE for the basic Unet at an ACR $= 8$ on *HELA-2* (note that due to the large amount of trained models required, *i.e*., 1190 trainings for all splits and methods, we mainly evaluated scenarios with two annotation types on *HELA-2*). The multi-task Unet with a classification branch and the pseudo-label filtering augmented methods Pseudo-label and Con2R lead to better results at $4 <$ ACR < 32 but are equally unsuited for low-annotation regimes. In contrast, FixMatch and *DSP* lead to significantly better results for all ACRs. Interestingly, FixMatch and DSP even improve in the fully-supervised case ($ACR = 1$).

Supervision $\mathcal{M} + \mathcal{B}$: Results for the scenario with pixel-masks and bounding boxes are given in Fig. 3b. We see that all approaches which exploit boxes can improve over the baseline for $ACR > 2$. Interestingly, *DSP* leads to very strong results, for large ACRs even better than FixMatch ($+3.8\%$ and $+4.8\%$ DICE over FixMatch at ACRs 32, 64).

Supervision $\mathcal{M} + \mathcal{P}$: Results for learning from masks and points are given in Fig. 3c. Again, all methods perform better than a naive baseline for $ACR > 2$. The slowest decline in DICE over ARCs is again obtained by *DSP*: $+1.3\%$, $+2.8\%$, $+5.2\%$, $+5.1\%$ for ACRs $8 - 64$. In comparison with Fig. 3b and Fig. 3a, we find a confirmation for the intuition that boxes provide more information than points, which again provide more information than image labels.

Supervision $\mathcal{M} + \mathcal{I} + \mathcal{B} + \mathcal{P} + \mathcal{U}$: Results from experiments with diverse annotation types are obtained on all datasets for the naive baseline and the best two techniques from the previous analysis, *i.e*., FixMatch and DSP, resulting in 455 model trainings. Results on *HELA-2* are given in Fig. 3d. We see that an expert-centric training algorithm can pay off: the accuracy with DSP declines only slightly with fewer and fewer masks. At an ACR $= 64$ with merely 1.6% pixel-wise masks (less than 40), the modelling and training as Decoupled Semantic Prototypes still leads to 49.5% DICE, which is comparable to the basic Unet under

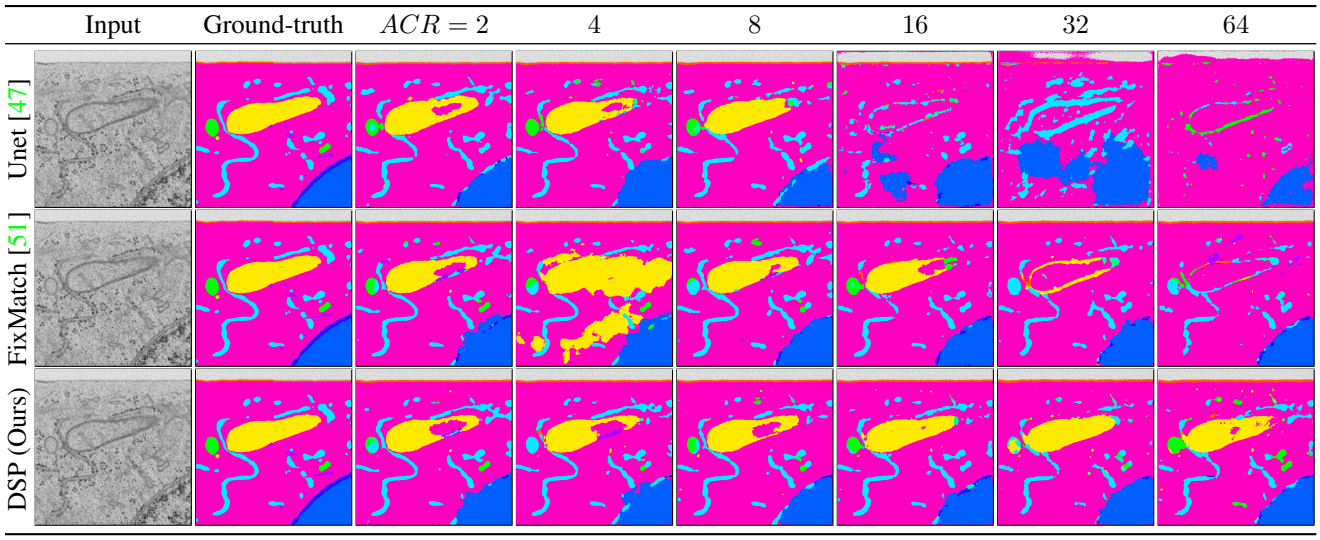

Figure 4. Qualitative results on *HELA-2*. From left to right: test image, ground-truth segmentation of the organelle structures, and predicted segmentations from models trained with diverse annotation types and decreasing number of annotated pixel-wise masks.

full supervision with 50.1% DICE. Compared to FixMatch, at ACR $= 64$, DSP can improve segmentation quality by 12.8% total DICE. We were surprised to find that with this style of diverse annotations, DSP performs better than in any scenario with annotation type pairs. This may indeed be due to the mix of annotations, which may enable a correct identification of central points and simultaneously allow learning of spatial extent from boxes.

Additional results on the remaining cell datasets are given in Figures 3e to 3g. We again observe improved results from *DSP* over FixMatch, and in turn from FixMatch over the Basic Unet. An exception is *JURKAT-1*, which proves to be challenging.

Hyperparameter sensitivity studies To investigate effects of hyperparameters of *DSP*, we report results from ablation studies on the *HELA-2* mixed, ACR $= 8$ scenario in Table 1. In the first part, we see that adding L_{DSP} to the simple semi-weak baseline L_{PLF} gives a clear improvement, confirming that the annotation type-based losses indeed help. Adjusting the weights for different annotation types to result in similar magnitudes did not help (3rd row). Too low factors also decreased results (5th row). We settle for a simple $\lambda_l = 0.1$ weighting. In the second part, we see that the temperature τ is an important factor and we observe best results for $\tau = 0.05$. We further conclude from part three that the optimal number of prototypes per class is $|P_c| = 5$, although we expect that this depends on dataset aspects and intra-class variances. Finally, a pure supervised training with only L_{CE} is sub-optimal as seen in part four.

5.4. Qualitative results

For a final analysis, we can visually inspect differences of a Unet baseline, FixMatch, and DSP in Figure 4. While the baseline already struggles with small ACRs, FixMatch still leads to visually decent results up to ACR 16. Interestingly, DSP is able to perform organelle segmentation even at ACRs of 32 or 64, which underlines the benefit of expert-centric and annotation-efficient training algorithms.

6. Conclusion

Our paper focused on expert-driven domains, where exact annotations are hard to obtain and one is usually confronted with different types of coarse annotations. To tackle this problem, we introduced *Decoupled Semantic Prototypes* by posing semantic segmentation as a novel class-specific prototype association problem that can be easily used for integrating different types of annotation. The prototypes can be trained with a Decoupled Contrastive Loss which we adapted to scenarios with different annotation types. We studied the benefits of our method for electron microscopy cell organelle segmentation and investigated how our models and several baselines perform when provided with smaller quantities of pixel-wise annotations. In these scenarios, our method is able to significantly slow down the degradation in accuracy with reduced annotations. This allows for a new level of flexibility in annotation processes and focuses on what experts can offer rather than forcing them to obey to restrictions of a learning method.

Limitations Our formulation of ACR does not yet consider different costs for different types of weak annotations. Instead, it solely focuses on substituting base annotations with any weaker type. This is sub-optimal, and it will be interesting to consider a holistic measure which respects individually specified costs in future work.

Acknowledgement This work was performed on the HoreKa supercomputer funded by the Ministry of Science, Research and the Arts Baden-Württemberg and by the Federal Ministry of Education and Research.

References

[1] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4981–4990, 2018. 2

[2] Inigo Alonso, Alberto Sabater, David Ferstl, Luis Montesano, and Ana C Murillo. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8219–8228, 2021. 2

[3] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *European Conference on Computer Vision*, pages 456–473. Springer, 2022. 3

[4] Wonho Bae, Junhyug Noh, Milad Jalali Asadabadi, and Danica J Sutherland. One weird trick to improve your semi-weakly supervised semantic segmentation model. *arXiv preprint arXiv:2205.01233*, 2022. 2

[5] 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*, pages 549–565. Springer, 2016. 2

[6] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material recognition in the wild with the materials in context database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3479–3487, 2015. 2

[7] Rodrigo Benenson, Stefan Popov, and Vittorio Ferrari. Large-scale interactive object segmentation with human annotators. In *CVPR*, 2019. 1

[8] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019. 5

[9] Davis Blalock, Jose Javier Gonzalez Ortiz, Jonathan Frankle, and John Guttag. What is the state of neural network pruning? *Proceedings of machine learning and systems*, 2:129–146, 2020. 5

[10] Hongjun Chen, Jinbao Wang, Hong Cai Chen, Xiantong Zhen, Feng Zheng, Rongrong Ji, and Ling Shao. Seminar learning for click-level weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6920–6929, 2021. 2

[11] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 1

[12] Junsuk Choe, Seong Joon Oh, Seungho Lee, Sanghyuk Chun, Zeynep Akata, and Hyunjung Shim. Evaluating weakly supervised object localization methods right. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3133–3142, 2020. 2

[13] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1, 2

[14] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1635–1643, 2015. 2

[15] 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*, pages 248–255. Ieee, 2009. 2

[16] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 7

[17] Jose Dolz, Christian Desrosiers, and Ismail Ben Ayed. Teach me to segment with mixed supervision: Confident students become masters. In *International Conference on Information Processing in Medical Imaging*, pages 517–529. Springer, 2021. 2

[18] Rosana El Jurdi, Caroline Petitjean, Paul Honeine, and Fahed Abdallah. Bb-unet: U-net with bounding box prior. *IEEE Journal of Selected Topics in Signal Processing*, 14(6):1189–1198, 2020. 2

[19] Mark Everingham, SM Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 1

[20] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59(2):167–181, 2004. 2

[21] Feng Gao, Minhao Hu, Min-Er Zhong, Shixiang Feng, Xuwei Tian, Xiaochun Meng, Zeping Huang, Minyi Lv, Tao Song, Xiaofan Zhang, et al. Segmentation only uses sparse annotations: Unified weakly and semi-supervised learning in medical images. *Medical Image Analysis*, page 102515, 2022. 2

[22] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010. 6

[23] Matthew D Guay, Zeyad AS Emam, Adam B Anderson, Maria A Aronova, Irina D Pokrovskaya, Brian Storrie, and Richard D Leapman. Dense cellular segmentation for em using 2d–3d neural network ensembles. *Scientific reports*, 11(1):1–11, 2021. 1

[24] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. *arXiv preprint arXiv:2209.08575*, 2022. 1

[25] Larissa Heinrich, Davis Bennett, David Ackerman, Woohyun Park, John Bogovic, Nils Eckstein, Alyson Petruncio, Jody Clements, Song Pang, C Shan Xu, et al.

Whole-cell organelle segmentation in volume electron microscopy. *Nature*, 599(7883):141–146, 2021. 5

[26] 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*, pages 7014–7023, 2018. 2

[27] Mostafa S Ibrahim, Arash Vahdat, Mani Ranjbar, and William G Macready. Semi-supervised semantic image segmentation with self-correcting networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12715–12725, 2020. 2

[28] 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*, pages 876–885, 2017. 2

[29] Viveka Kulharia, Siddhartha Chandra, Amit Agrawal, Philip Torr, and Ambrish Tyagi. Box2seg: Attention weighted loss and discriminative feature learning for weakly supervised segmentation. In *European Conference on Computer Vision*, pages 290–308. Springer, 2020. 2

[30] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013. 2, 5, 7

[31] Hyeonsoo Lee and Won-Ki Jeong. Scribble2label: Scribble-supervised cell segmentation via self-generating pseudo-labels with consistency. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 14–23. Springer, 2020. 2

[32] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5267–5276, 2019. 2

[33] Junnan Li, Caiming Xiong, and Steven CH Hoi. Comatch: Semi-supervised learning with contrastive graph regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9475–9484, 2021. 5

[34] Qizhu Li, Anurag Arnab, and Philip HS Torr. Weakly-and semi-supervised panoptic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 102–118, 2018. 2

[35] Yanwei Li, Hengshuang Zhao, Xiaojuan Qi, Yukang Chen, Lu Qi, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Fully convolutional networks for panoptic segmentation with point-based supervision. *arXiv preprint arXiv:2108.07682*, 2021. 2

[36] 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*, pages 3159–3167, 2016. 2

[37] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 2

[38] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 6

[39] Wenfeng Luo and Meng Yang. Semi-supervised semantic segmentation via strong-weak dual-branch network. In *European Conference on Computer Vision*, pages 784–800. Springer, 2020. 2

[40] Pawel Mlynarski, Hervé Delingette, Antonio Criminisi, and Nicholas Ayache. Deep learning with mixed supervision for brain tumor segmentation. *Journal of Medical Imaging*, 6(3):034002, 2019. 2, 7

[41] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 4990–4999, 2017. 1

[42] 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*, pages 1742–1750, 2015. 2

[43] Deepak Pathak, Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional multi-class multiple instance learning. *arXiv preprint arXiv:1412.7144*, 2014. 2

[44] Simon Reiß, Constantin Seibold, Alexander Freytag, Erik Rodner, and Rainer Stiefelhagen. Every annotation counts: Multi-label deep supervision for medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9532–9542, 2021. 2

[45] Simon Reiß, Constantin Seibold, Alexander Freytag, Erik Rodner, and Rainer Stiefelhagen. Graph-constrained contrastive regularization for semi-weakly volumetric segmentation. In *European Conference on Computer Vision*, pages 401–419. Springer, 2022. 2, 5, 7

[46] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Alexander G Schwing, and Jan Kautz. Ufo²: A unified framework towards omni-supervised object detection. In *European Conference on Computer Vision*, pages 288–313. Springer, 2020. 2

[47] 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*, pages 234–241. Springer, 2015. 6, 7, 8

[48] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3):309–314, 2004. 2

[49] Constantin Seibold, Jens Kleesiek, Heinz-Peter Schlemmer, and Rainer Stiefelhagen. Self-guided multiple instance learning for weakly supervised thoracic diseaseclassification and localizationin chest radiographs. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 2

[50] Constantin Marc Seibold, Simon Reiß, Jens Kleesiek, and Rainer Stiefelhagen. Reference-guided pseudo-label generation for medical semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2171–2179, 2022. 2

[51] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 2, 5, 7, 8

[52] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang 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

[53] Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. *arXiv preprint arXiv:2206.14486*, 2022. 5

[54] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021. 3

[55] Liyan Sun, Jianxiong Wu, Xinghao Ding, Yue Huang, Zhong Chen, Guisheng Wang, and Yizhou Yu. A teacher-student framework for liver and tumor segmentation under mixed supervision from abdominal ct scans. *Neural Computing and Applications*, pages 1–15, 2022. 2

[56] Meng Tang, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov, and Christopher Schroers. Normalized cut loss for weakly-supervised cnn segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1818–1827, 2018. 2

[57] Meng Tang, Federico Perazzi, Abdelaziz Djelouah, Ismail Ben Ayed, Christopher Schroers, and Yuri Boykov. On regularized losses for weakly-supervised cnn segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 507–522, 2018. 2

[58] Zhi Tian, Chunhua Shen, Xinlong Wang, and Hao Chen. Boxinst: High-performance instance segmentation with box annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5443–5452, 2021. 5, 7

[59] Gabriele Valvano, Andrea Leo, and Sotirios A Tsaftaris. Learning to segment from scribbles using multi-scale adversarial attention gates. *IEEE Transactions on Medical Imaging*, 40(8):1990–2001, 2021. 2

[60] Bin Wang, Guojun Qi, Sheng Tang, Tianzhu Zhang, Yunchao Wei, Linghui Li, and Yongdong Zhang. Boundary perception guidance: A scribble-supervised semantic segmentation approach. In *IJCAI International joint conference on artificial intelligence*, 2019. 2

[61] Guotai Wang, Maria A Zuluaga, Wenqi Li, Rosalind Pratt, Premal A Patel, Michael Aertsen, Tom Doel, Anna L David, Jan Deprest, Sébastien Ourselin, et al. Deepigeos: a deep interactive geodesic framework for medical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1559–1572, 2018. 7

[62] Pei Wang, Zhaowei Cai, Hao Yang, Gurumurthy Swaminathan, Nuno Vasconcelos, Bernt Schiele, and Stefano Soatto. Omni-detr: Omni-supervised object detection with transformers. *arXiv preprint arXiv:2203.16089*, 2022. 2

[63] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12275–12284, 2020. 2

[64] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 1

[65] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16684–16693, 2021. 2

[66] Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun. Decoupled contrastive learning. *arXiv preprint arXiv:2110.06848*, 2021. 2, 3

[67] Yu Zeng, Yunzhi Zhuge, Huchuan Lu, and Lihe Zhang. Joint learning of saliency detection and weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7223–7233, 2019. 2

[68] Ke Zhang and Xiahai Zhuang. Cyclemix: A holistic strategy for medical image segmentation from scribble supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11656–11665, 2022. 2

[69] Xiangyun Zhao, Raviteja Vemulapalli, Philip Andrew Mansfield, Boqing Gong, Bradley Green, Lior Shapira, and Ying Wu. Contrastive learning for label efficient semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10623–10633, 2021. 2

[70] Zhuo Zhao, Lin Yang, Hao Zheng, Ian H Guldner, Siyuan Zhang, and Danny Z Chen. Deep learning based instance segmentation in 3d biomedical images using weak annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 352–360. Springer, 2018. 2

[71] Yuanyi Zhong, Bodi Yuan, Hong Wu, Zhiqiang Yuan, Jian Peng, and Yu-Xiong Wang. Pixel contrastive-consistent semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7273–7282, 2021. 2

[72] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 2

[73] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 1

[74] Tianfei Zhou, Wenguan Wang, Ender Konukoglu, and Luc Van Gool. Rethinking semantic segmentation: A prototype view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2582–2593, 2022. 3

[75] Yanning Zhou, Hang Xu, Wei Zhang, Bin Gao, and Pheng-Ann Heng. C3-semiseg: Contrastive semi-supervised segmentation via cross-set learning and dynamic class-balancing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7036–7045, 2021. 2

[76] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking pre-training and self-training. *Advances in neural information processing systems*, 33:3833–3845, 2020. 2