Weakly-Supervised Semantic Segmentation

Research Proficiency Exam

David Paredes

Supervisor: Dr. Dimitris Samaras

Department of Computer Science

State University of New York at Stony Brook

September 2020

Abstract

Over the past decade, deep learning has revolutionized the computer vision field. In particular, image segmentation methods have shown big improvements in typical fully-supervised learning settings. Unfortunately, these methods required datasets of costly segmentation annotations. For example, fine annotations at pixel level of an image from the CityScapes dataset required an average time of 90 minutes. Weakly-supervised segmentation methods aim to reduce the need of labor-intensive annotated datasets by training the models with weaker forms of supervision such as bounding boxes or point annotations.

In this report, we explore existing approaches that try to leverage weak supervision labels for the task of semantic segmentation. Then, we propose a new approach with point annotations for semantic segmentation of immune cells in multiplex immunohistochemistry imagery. We show how our method can achieve high quality segmentation details.

Contents

1 Introduction

The field of computer vision has a generic goal of extracting semantic information from digital images or videos, which represent the visual world. Segmentation is a core task for computer vision that empowers a wide range of applications such as autonomous navigation, medical imaging, satellite imagery, etc. See Fig. 1 for examples.

The task of segmentation breaks images or videos into appealing entities that can enable the task of recognition, which is useful for additional tasks. Despite the benefits shown by deep learning methods in segmentation tasks, a clear drawback of these supervised learning approaches is that they rely heavily on large datasets of pixel-wise labeled images. For instance, they contain detailed examples of region boundaries of objects such as cat, dots, etc. However, from a point of view of human vision, we can argue that humans actually learn from far less supervision. For instance, a mother might point at a toy and her child can internally do the mapping of the object in the visual scene. In this example, we can say that weak supervision works in the context that you have the ability to create objects.

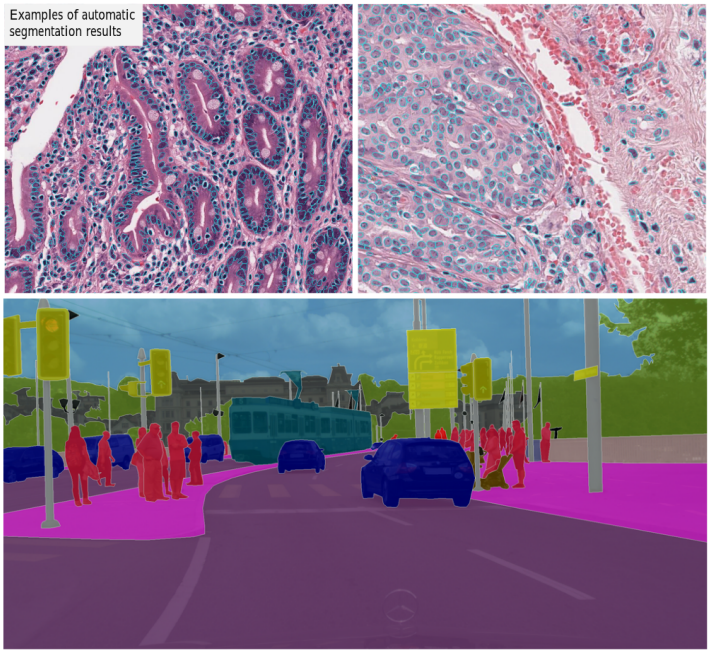


Figure 1: Datasets for image segmentation. Top: Sample image of a nuclei segmentation dataset for medical imaging. Bottom: Sample image of a autonomous navigation dataset

As a consequence, being able to learn from weak sources of supervision is one of the fundamentals abilities of how humans learn. For this reason, designing algorithms for weakly supervised problems would be beneficial to scale up computer vision tasks.

In Section 2, we cover the basics related to image segmentation in general. In Section 3, we review the most important types of weak supervision that has been applied in the literature of weakly-supervised semantic segmentation. Additionally, it contains related applications such as object detection. In Section 4, we propose a novel approach for weakly-supervised cell segmentation. We have only access to point annotations during training and we extend this weak labels into superpixels as pseudo masks for training a fully convolutional network. Moreover, we apply multi-resolution losses at the intermediate feature maps. The predictions of out final network are combined with a color deconvolution method to further improve the segmentation details of the cells. In the last Section 5, we argue how important our proposed method can be used for future applications. Moreover, we present ideas for future work to further reduce the gap between weakly supervised problems an fully supervised methods for segmentation.

2 Background

Image segmentation is a computer vision task, whose goal is to partition an input image into segments in order to reduce the complexity of the image. The result of image segmentation can be considered as a step closer towards reasoning about the content on images. Moreover, segmentation can further be applied on videos.

2.1 Types of Image Segmentation

2.1.1 Semantic segmentation

Semantic segmentation predicts one label for each pixel in an image from a defined set of classes. It is important to point out that if there multiple countable objects of the same label, the predicted segmentation map does not distinguish the objects since we only care about the category at the pixel level. Fig. 2 depicts an example of semantic segmentation from the training set of the Pascal VOC 2012 dataset.

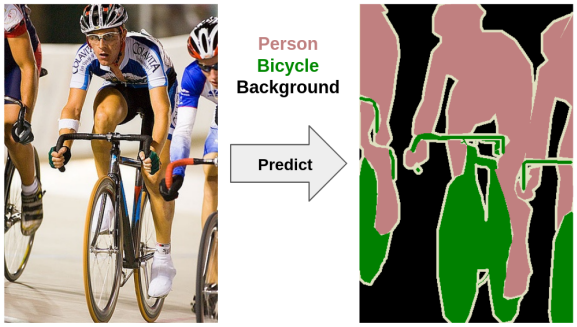


Figure 2: Pascal VOC 2012 semantic segmentation task

2.1.2 Instance segmentation

Instance segmentation corresponds to the task of identifying all pixels belonging to every object of interest in an image. It can be thought of an object detection problem while also outlining all objects of interest. Unlike semantic segmentation, the predicted segmentation map does not assign a class to the background pixels. In the Fig. 3, the outputs shows 3 objects detected from the training set of the MS COCO dataset.

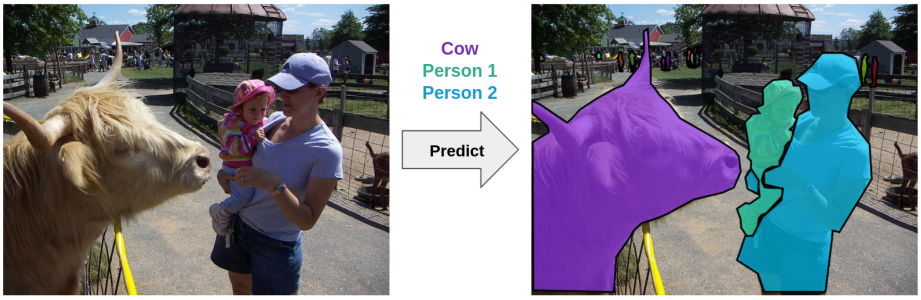


Figure 3: MS COCO instance segmentation task

2.1.3 Panoptic segmentation

Kirillov et al. [25] introduced panoptic segmentation task to unify both semantic and instance segmentation. They describe the content of an image as *stuff* and *things* classes. The class *stuff* is studied as uncountable, amorphous regions of similar material or texture such as road, grass, sky. This is usually formulated for the task of semantic segmentation. On the contrary, the class *things* corresponds to countable objects such as people, animals, cars. This is naturally formulated for the task of instance segmentation. Formally, the panoptic segmentation task predicts a segmentation map that assigns two labels for each pixel. One is a semantic label and the other one is an instance id. In the case where a pixel refers to the *stuff* class, the instance id is ignored. Fig. 4 depicts an example of panoptic segmentation.

2.1.4 Others

Scene parsing. This task was introduced as pixel-wise dense labeling semantic segmentation in complex scenes and it is characterized for using a high number of classes. A representative benchmark is ADE20K dataset [67, 68] and it covers 150 semantic categories. See Fig. 4 for an example of the complexity of the scenes.

Object co-segmentation. Rother et al. [50] is the first work to propose *co-segmentation* as segmenting common parts of a pair of images. From recent works, this segmentation is approached as segmenting common objects from a set of images. The idea is that segmenting a set of images jointly can achieve better performance than segmenting the images independently as shown in [10].

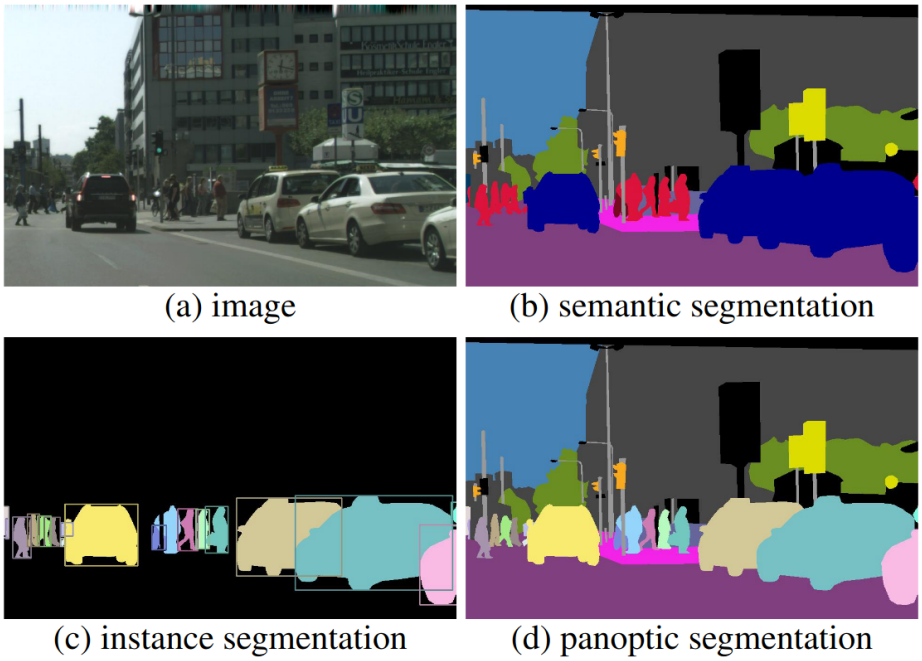


Figure 4: Examples of differences among semantic, instance and panoptic segmentation. Figure from [25]

2.2 Image segmentation evaluation

2.2.1 Intersection over Union

The Intersection over Union (IoU), also known as Jaccard index, is a metric that measures the agreement (overlapping) of the predicted segmentation map and the ground truth segmentation map. This is the standard metric for semantic segmentation that evaluates each class and it is defined as follows

$$IoU = \frac{\text{Area of overlap}}{\text{Area of union}} \tag{2.1}$$

The range of values are between 0 an 1, where closer to 1 means better performance. Since a dataset has multiple classes, $mIoU$ is defined as the average of IoU for all the classes.

2.2.2 Average Precision

The Average Precision (AP) is the standard metric for instance segmentation. Before defining AP, it is important to introduce True Positive (TP), False Positive (FP) and False Negative (FN) in this context. A TP corresponds when the IoU between the predicted

Figure 5: Illustration of complex scenes in ADE20K dataset [67, 68] for scene parsing. Top row are samples images and bottom row are their segmentation maps.

Figure 6: Example of object co-segmentation for the class of cars from [10]

and the ground truth exceeds the predefined threshold. A FP corresponds when a predicted instance cannot be associated to a ground truth instance because their IoU is lower than predefined threshold. A FN corresponds when a ground truth instance cannot be associated to a predicted instance because their IoU is lower than predefined threshold. For a given threshold, we can define the precision and recall as follows

$$Precision = \frac{TP}{TP + FP} \tag{2.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.3}$$

It means that we can plot precision-recall curve for multiple IoU thresholds. Formally, we define the AP as the area the curve for a specific range of thresholds. The primary metric of the MS COCO dataset [32] is defined as the average over 10 IoU thresholds from 0.5 until 0.95 every 0.05.

2.2.3 Panoptic Quality metric

Panoptic Quality (PQ) is a standard metric for panoptic segmentation. The idea is to unify the metrics for semantic and instance segmentation an also to treat all the categories equally. For each class, we compute two steps: segment matching and PQ computation. Consider the example of Fig. 7, the ground truth segments and the predicted segments have matched if their IoU is at least 0.5. It is a unique matching because of the non-overlapping property. Therefore, one predicted segment corresponds at most one ground truth segment.

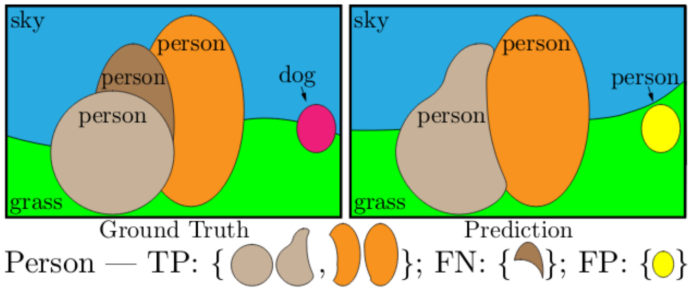


Figure 7: Toy example of predicted segmentation and ground truth segmentation for the task of panoptic segmentation. Segments of the same color are matched because their IoU is larger than 0.5.

Formally, for a predicted segment p and a ground truth segment g, the calculation of PQ is defined as follows

$$\mathrm{PQ} = \frac{\sum_{(p,g)\in TP} \mathrm{IoU}(p,g)}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|} \tag{2.4}$$

or it can be rewritten as *segmentation quality*(SQ) term and a *recognition quality*(RQ) term

$$\mathrm{PQ} = \underbrace{\frac{\sum_{(p,g)\in TP} \mathrm{loU}(p,g)}{|TP|}}_{\text{segmentation quality (SQ)}} \times \underbrace{\frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}}_{\text{recognition quality (RQ)}} \tag{2.5}$$

It is important to mention that all segments are equally important regardless of their size. Similarly as in AP, the PQ for each class is calculated independently, then the final PQ values is the average over all the classes.

3 Weakly-Supervised Semantic Segmentation

For fully-supervised semantic segmentation, the desired output mask is provided for each input image at the pixel level. However, weakly-supervised settings for this task attempt to build models by learning from weak signals of supervision (annotations). Berman et al. [4] evaluated the time for different types annotations in the Pascal VOC 2012 dataset as shown in Fig. 8.

In the following subsections, we discuss the different types of weak supervision that has been proposed in the literature for the task of semantic segmentation.

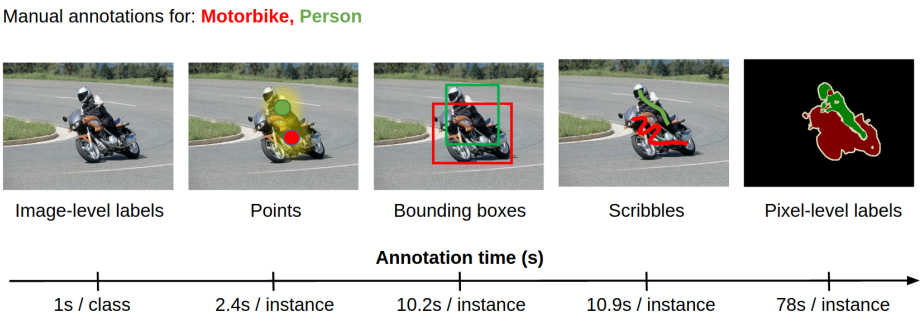


Figure 8: Estimated time for annotations for Pascal VOC 2012 dataset [4]

3.1 Types of weak supervision

3.1.1 Image-level labels

This is considered the simplest and most efficient annotations so it has become the most common type of weak supervision for semantic segmentation. The only annotations available for training are the image-level labels, which is a list of classes that are included in the image. See Fig. 8 for an example. From the literature, we can organize the popular approaches as follows:

- End-to-end Learning with Constraint Loss:
  - Multiple Instance Learning:
    In this context, the framework of multiple instance learning means that each image given a class label, contains at least one pixel that also match the class label. [41] introduces a constrain in the loss to put more weight on important pixels

for classification. The key is to aggregate the intermediate pixel-level scores into a single image-level classification score, which has to be maximized. Similarly, [40] operates only over the max scoring pixel in the coarse intermediate heatmaps and backpropagates through the network.

The benefits of these formulations is that it can be trained end-to-end. The final segmentation is obtained from the top class predicted of the coarse heatmap and interpolated to the original image resolution.

- Constrained Convolutional Neural Networks:
  [39] improves the multiple instance learning framework by enforcing some constrains such as foreground, background, size and suppression constrains. At inference time, they apply a fully connected conditional random field as a refinement step [27].

- Pseudo Pixel-level Masks:
  - Additional data.
    These cases are also known as webly supervised methods because the external data is collected from the web.

    One of the first works is [60], which extract two sets to distinguish foreground and background from Flickr. They used as prior the saliency map of an object detector as pseudo masks. Then another network is trained on the pseudo masks of the two sets. [22] collects three external dataset from the web. The conditions of the datasets are: images with white background, images with common background scenes and realistic images of each object category. The pseudo masks of these datasets are used in a three-stage training pipeline and refinement step. [18] uses additional web-crawled videos in such a way that no additional supervision is required. After learning from the weak labels, the model is used to eliminate irrelevant frames. The key idea is to learn a spatio-temporal segmentation of objects by a graph formulation using superpixels. Then, it continues learning the segments from videos at the frame level.
  - Object proposals.
    [43] proposes an object localization branch to classify regions proposals generated with selective search. It aggregates the proposals to generate segmentation masks as pseudo masks to train the semantic segmentation mask. [15]

uses in the first stage a neural attention heatmap and an object proposal heatmap. Both are use to generate object instances. In the second stage, it trains an instance classifier for instance filtering as well as a triplet loss based on metric learning for outlier detection. In the last stage, it combines the previous output with the initial attention maps and object heatmap into a final probability map.

– Class Activation Mapping.
Zhou et al. [66] introduced Class Activation Mapping (CAM) as an attention mechanism that identifies the most discriminative area using a pre-trained classification method. The idea is to replace the fully connected layers by a Global Average Pooling (GAP) layer, followed by a one-by-one convolutional layer. For a specific class of interest, these two modifications introduced a weight sum of the class activation maps as shown in Fig. 9

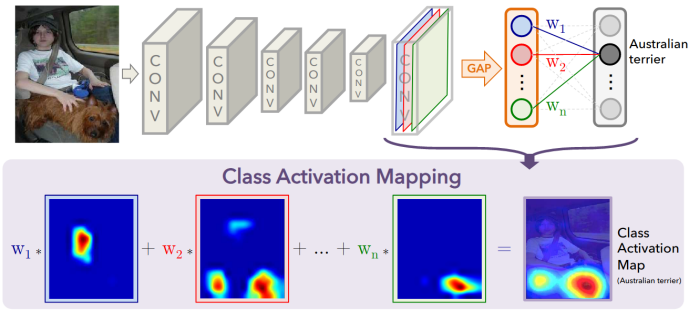


Figure 9: CAM localise the most discriminative regions for a specific class [66].

The most popular methods using CAM are the following:

* Seed and expand.
CAM produces high-precision weak localisation cues (regions). The idea of [26] is to expand these cues to align with the object boundaries. This is done by penalizing discontinuous segmentations with a fully connected CRF loss. Huang et al. [21] used a traditional image segmentation method of seeded region growing to generate the pseudo masks for training.
* Adversarial Erasing.
[59] introduces the idea of *erasing* a small, yet the most discriminative region generated by CAM from an input image. Then, the classification network is retrained with the modified images in order to keep discovering

discriminative object regions. [19] tried to avoid mining unexpected background regions by using saliency maps as prior to identify background regions. There is a strategy of self-erasing in two decoder branches. The first branch erases the high peak CAM responses and it is the input for the second branch. [7] also combines saliency map (saliency detector) and CAM in a hierarchical approach without retraining the decoder. The image is erased in each step and the foreground mask is accumulated.

* Exploiting Saliency.
  Saliency maps can be used as prior knowledge. [35] uses CAM to find areas of object with high confidence. A saliency detector is used to find the object mask that corresponds to the CAM output area. These pseudo masks are used for training a segmentation model. [65] proposes to jointly training semantic segmentation and saliency detection can improve the segmentation mask.
* Multi-dilated Convolution.
  [61] is a simple approach that uses standard classification networks with multiple dilated convolutional blocks. They found that different dilation rates can enlarge the receptive field and improve the localization of discriminative areas using CAM.

3.1.2 Video-class labels

Similarly as image-level labels, [54] uses weakly-annotated videos at the frame level. Ideally, the motion cues can help to detect better object boundaries based on the "common fate", one of the Gestalt laws of grouping. The approach of [54] is to exploit the motion cues of videos as soft constraints. They used an object segmentation model to segment the motion from the videos. One of the assumptions is that each video is limited to a single category label.

3.1.3 Object bounding boxes

Annotation of bounding boxes is also one of the most popular types of weak supervision. The annotations consist of the class object label as well as the position of a bounding box of the object. An example of this annotation can be seen in Fig. 8. One of the most seminal works in this category for segmentation is GrabCut [49]. Typical approaches are the following:

- Pseudo Pixel-level Masks:
  [24] explore how robust are convolutional neural networks to noisy segmentation masks. They mask the whole box as a pseudo mask for the object and they show that the predicted segmentation mask improves with more iterations. This work is considered an important baseline. Similarly, BoxSup [11] recursively train a convolutional neural network with segment object proposals. The pseudo mask gets updated with the updated network.

- Cut-and-Paste:
  [46] introduces the idea of cut-and-paste object in a scene. The key is that if you are able to segment an object, therefore you can cut out the object and place it in another scene. This fake image is trained in an adversarial setting until the object mask can fool the discriminator by producing more realistic images.

3.1.4 Object extreme points

A bounding box for an object is defined by four points, and these points are actually outside the object. Instead, extreme points are defined as the left-most, right-most, top and bottom pixels inside the object. These points defined more information of the object since they are exactly on the object boundary. Therefore, a tight bounding box can easily be drawn as shown in Fig. 10. This weak supervision has been explored in related works such as [34, 37] and it is useful for guided segmentation (grabcut style). [34] creates a heatmap using a 2D gaussian blur on each extreme point and concatenate this heatmap to the inpute image. Then is trained with weighted cross-entropy with ResNet101 as backbone network.

3.1.5 Point-level supervision

The annotator select a point per object in the image and its class object label. Even though this type of annotations can be considered the least informative, they can dramatically alleviate the time of annotations as shown in Fig. 8. We can compare this weak supervision to the way how humans refer to objects by pointing as a mechanism that supports the communication.

Works on the domain of semantic segmentation with point supervision are [4, 44, 45, 58]. [4] uses an image-level supervision loss and a point level-supervision loss. They key idea is to used an objectness prior that provides a probability for each pixel if they belong to an object class from a pre-trained network. [44] proposes a novel

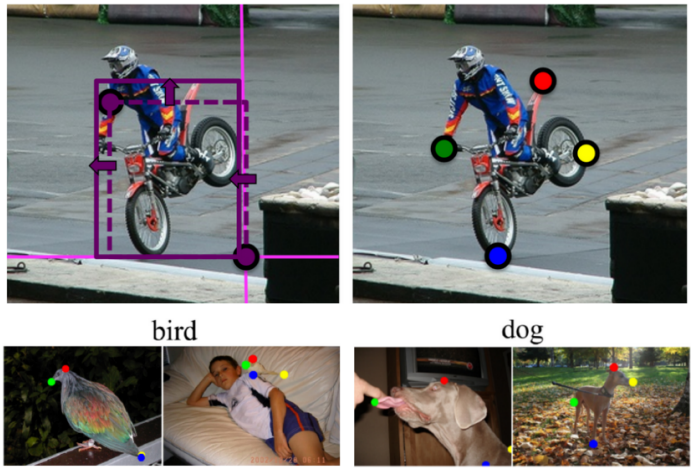


Figure 10: Top left shows how tight is a bounding box from extreme points. Top right and bottom row are examples of extreme points annotations of instances [37].

deep metric learning on embedding features of intra-category and inter-category. Points of the same category have more similar representations than those of different categories. [45] is the first weakly supervised method that incorporates dot annotations for nuclei segmentation in H&E stained histopathology images. They assumed that most nuclei shapes are convex regions, so segmentation masks are generated with Voronoi diagrams. After the segmentation network is trained, they used a fine-tuning step with a dense conditional random field loss. Moreover, there are other applications such as object counting [28], instance segmentation [29] and object detection [38].

3.1.6 Scribbles

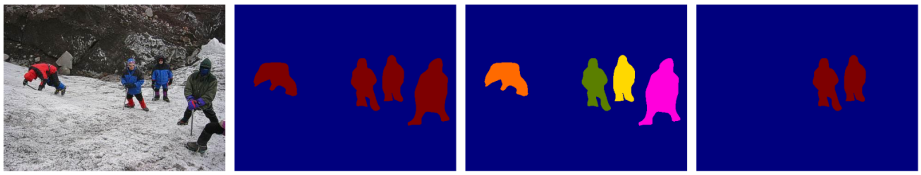
They can be considered as the generalization of point-level supervision. As shown in Fig. 8, only a subset of pixels inside the object are labeled with the object class. This annotations are also known as free-form squiggles in the literature.

Relevant works using scribbles as supervision for semantic segmentation are [4, 31, 62, 6, 57, 52]. [31] has two components. The fully convolutional network provides semantic predictions for the graphical model, which propagates the labels for network learning in a two-step iterative optimization. The graph defines unary and pairwise terms based on a superpixel segmentation. The unary potential term depends on whether the scribble is inside the superpixel. The pairwise potential term calculates the color similarity and texture

similarity among the superpixels. [52] introduces a new loss based on CRF that evaluates the classical *normalized cut* for unlabeled pixels and partial cross entropy on labeled pixels. [57] uses DeepLab as a backbone network and adds two branches. The first one refine segmentation results by combining high and low level features with a partial cross-entropy loss from [52]. The second branch extract class-agnostic edge features to guide the segmentation to localize boundaries.

3.1.7 Image captions

It can be considered an advanced version of image-level labels, where we have natural language expressions to describe the image. This novel problem for segmentation was first introduced in [20]. Fig. 11 shows an example of using image captions as weak supervision for segmentation. A LSTM encodes the expression into a vector representation, a fully convolutional network extracts a spatial feature map from the image. These two outputs are fed into a fully convolutional classification network and upsampling network to predict the segmentation map. Additionally, [47] shows a very similar application but instead for the task of object localization.



(a) input image (b) object class segmentation of class ***people*** (c) object instance segmentation of class ***people*** (d) segmentation from expression *"people in blue coat"*

Figure 11: When image captions are used as weak supervision (d), it outputs a different segmentation mask than traditional image segmentation (b) and (c).

3.1.8 Eye gaze

How humans process visual scenes are reflected by the eye movements. Studies have confirmed that these eye movements are not only determined by the content of an image, but also biased towards tasks such as visual search. Related works are on the domain of object detection such as [3, 23, 36, 63]. For example in [3] as shown in Fig. 12, the annotators are asked to describe the scene in videos with language expressions. As expected, they instinctively start looking at positions in the scene that matches the words and

the human gaze is recorded. This weak supervision provides soft cues of where the descriptions are in the scene. In [3] they collect a video dataset with their language sentences descriptions as well as the human gaze. They used object proposals and multiple LSTM for encoding local and global features of CNNs applied at the frame level. The gaze information is converted using a linear mapping of left and right eyes to estimate the gaze position in the image.

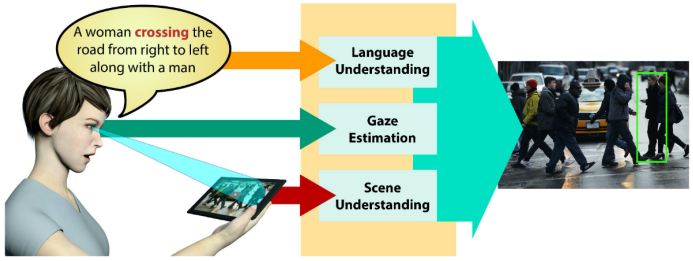


Figure 12: Eye movements are recorded while an annotator is describing the scene as shown in [3] for the task of object detection.

3.1.9 Multimodal image annotations

It is clear that humans can speak and point to describe their environment. Multimodal image annotations are a new attempt of connecting vision and language. Relevant work on this domain for the task of object location are [42, 56, 16]. In the recent work of [42], annotators are asked to describe and image with dense captions. At the same time, the voice and the hovering of the mouse inside the area described are recorded. The objective is to learn in order to learn audio-visual associations of the content of the scene. See Fig. 13 for an example of how these annotations look like. What's interesting from [42], is the generation of image captioning that can match the mouse trace. This controllable image captioning is a transformer-based image captioning model that ranks the most important object proposals.

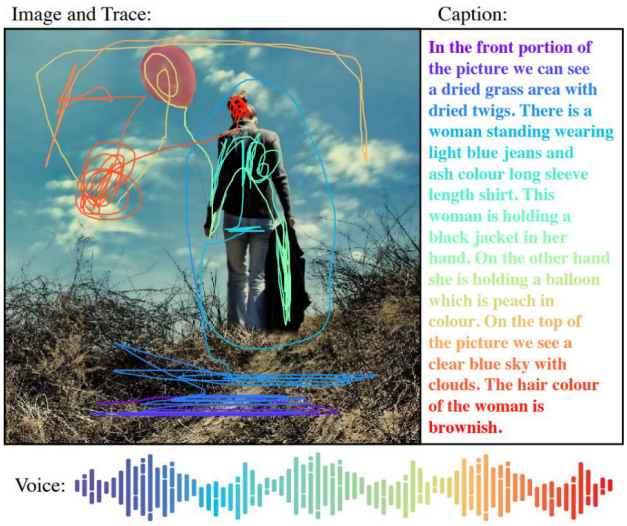


Figure 13: The caption, voice and mouse trace are simultaneously recorded. Each modality is represented by a color gradient. Image from [42].

4 Weakly-Supervised Cell Segmentation

In this section, we propose a novel approach for weakly-supervised semantic cell segmentation of multiplex immunohistochemistry images (mIHC). Our goal is to achieve high segmentation quality despite limited supervision and cell appearance heterogeneity in mIHC. In our approach, we first extend point annotations to mask annotations using superpixels. Second, we introduce multi-resolution supervision, i.e., comparing the prediction result at different layers/resolutions of the neural network. Finally, we observe that focusing on stains/colors can capture the fine-details of the cell near the boundary. Thus, we use a color deconvolution method (detecting masks corresponding to different colors/stains) to complement our segmentation network. Empirically, our method attains high quality segmentation results on sample patches from 4 whole slide images of pancreatic cancer tissue and we believe opens the way to future large-scale studies.

4.1 Problem description

Multiplex techniques allow the study of multiple cell types and the spatial relationships between them while maximizing the amount of information acquired from a single sample [5, 12, 17]. This is particularly important for the study of the tumor immune microenvironment, which has become an intense area of translational research focus. Multiplex immunohistochemistry (mIHC) and immunofluorescence (mIF) allow simultaneous labeling of 5 or more distinct cell types in the same tissue sample using colored chromogens or fluorophores, respectively. mIF tends to be more costly and requires a specialized microscope for image capture. On the other hand IHC is already routinely used in clinical medicine, and mIHC images can be captured by traditional bright field microscopy with a single, low-cost imaging step (10% of the cost of fluorescent staining), making mIHC the most rational choice for future large scale studies.

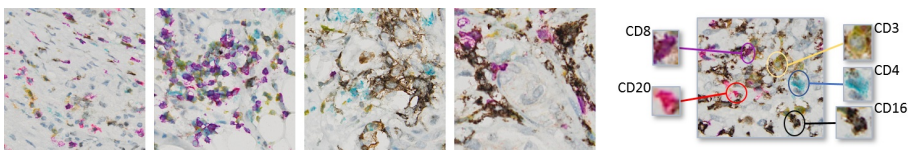


Figure 14: Left: Four sample images from mIHC whole slide images taken at the same scale and size illustrating the fuzzy cell boundaries and large variation in immune cell size and shape. Right: Stained immune cell types are magnified and labeled.

While scalable staining and image capture platforms are readily available for mIHC, automated image analysis platforms are lacking. Cell segmentation in mIHC images requires distinguishing between cell classes primarily based on color. Compared with other microscopic image modalities, there is high cell appearance heterogeneity in mIHC images. Cell shapes are highly variable; most lymphocytes are nearly circular with an average diameter of 8 micrometers (um), while macrophages range from discrete rounded cells of similar size, to elongated cells (∼20um) with projections extending in all directions. The edges of the cells may appear fuzzy due to chromogen properties, and boundaries between cells may be difficult to detect when cells are in close proximity. Furthermore, the appearance of the same cell types can be variable. Compared with H&E and fluorescence microscopy images, chromogenic staining has less clear delineation of the cell boundaries. Furthermore, nearby cells tend to be fused together and are hard to separate (see Fig. 14).

mIHC works by tagging an enzyme to a specific protein that is uniquely produced by a given cell type; the enzyme acts on the chromogen, producing a colored dye localized only at the cell type of interest. Individual cells of a given class may produce varying amounts of the specific proteins targeted by the mIHC stains, leading to differences in staining color intensities across a class. Furthermore, color spectra of different chromogen stains overlap significantly, making it difficult to distinguish between some classes. While all cell nuclei in the tissue are stained with hematoxylin (nuclear counterstain, blue), target proteins are often localized at the cytoplasm of the cell. In the case of lymphocytes, much of the cell volume is taken up by the nucleus, which introduces a significant co-staining issue. Therefore, a direct color deconvolution (unmixing different colors as cell masks) [8, 9, 33, 51, 53, 55] is both challenging and insufficient.

Another challenge in mIHC image segmentation is the lack of high quality manual annotations for training. Manual generation of highly detailed, high quality training data is challenging even for expert pathologists due to interactions between nearby biomarkers, co-existence of multiple stains in a cell, biomarker protein expression variation, and the large variation of macrophage shapes. There is a need for segmentation methods that can use weak annotations, for example, points placed at (approximately) the cell centers.

4.2 Methodology

We use weak annotation labels in the form of points placed in the centers of the cells. These point annotations are used to train and to validate our automatic segmentation method. We use superpixels to expand the labels of each point to its adjacent area, so that we have a per-pixel annotation to train our network. We compute superpixels in each input image using SLIC [2] (one of the commonly used superpixel methods). Each superpixel is assigned the same semantic label (i.e., the stain/color) as the point within it. This means all pixels within a superpixel are given the same semantic label. An empty superpixel is assigned the background label.

Fig. 15 illustrates the overview of our method. We first convert point annotations into a superpixel mask. Using this per-pixel annotation, we train a semantic segmentation network. We adopt the UNet [48] architecture. To address the large cell shape and size variations, we introduce multi-scale supervision, namely, provide loss-based supervision at different resolutions/layers of the decoder module. This way the network learns accurate representations at different resolutions. Details are provided in Section 4.2.1. Finally, since the superpixel mask itself is inaccurate in delineating cell boundaries, the segmentation model trained with the mask cannot be expected to produce correct details. To address this problem, we introduce a color decomposition network that is able to capture fine grain stain presence and composition. We then employ an ensemble method that combines the segmentation results of both networks. The final results enjoy the advantages of both networks and are of higher quality. See Section 4.2.2 for more details.

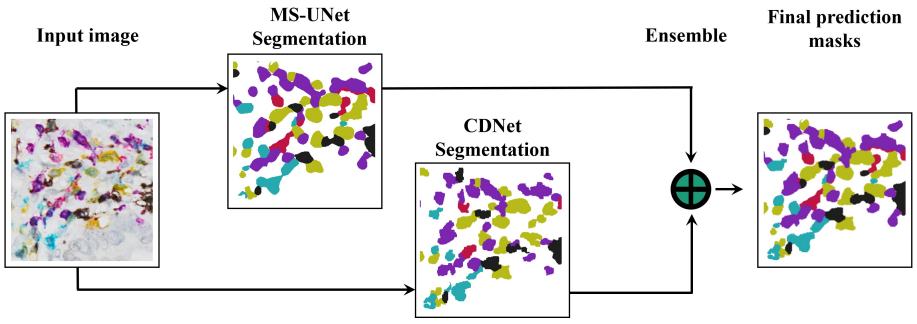


Figure 15: Pipeline of the ensemble method, which combines the predictions of the multi-resolution UNet (MS-UNet) and the color decomposition model (CDNet).

4.2.1 Segmentation Network

In this section, we propose a semantic segmentation network based on the UNet [48] architecture. The segmentation network has to deal with the fact that cells exhibit a large variation in shape and size (see Fig. 14 for an illustration). To address this challenge, we draw inspiration from how pathologists study tissue images. Pathologists often use multiple magnifications jointly so that they can take into consideration cell/tissue architectural features at different scales. By viewing images (especially mIHC images) at different scales, pathologists capture contextual information about shape and size variations.

Based on this insight, we introduce a multi-resolution supervision component to the segmentation network. In particular, we introduce additional supervision to different intermediate layers of the decoder module of UNet. The additional supervision enables the network to learn more discriminative features in those intermediate layers so that their feature representation better captures cells of different sizes/shapes. In the literature, this technique, called a deeply supervised network [30], has been applied in various domains [13, 14, 64, 69]. To the best of our knowledge, *we are the first to apply deep supervision to microscopic image analysis tasks.*

The architecture of our multi-resolution UNet (MS-UNet) is illustrated in Fig. 16. More specifically, we input the intermediate feature representation of each of the layers into a convolutional layer for semantic label prediction. The prediction is of the same resolution as the layer. We use a downsampled superpixel mask to supervise this prediction (via cross-entropy loss). In this manner, we can enforce the segmentation network to learn better representations at different resolutions, from fine to coarse. The total loss L can be formulated as follows

$$L = L_{CE}(Y, \widehat{Y}) + \sum_{i=1}^{n} \lambda_i L_{CE}(Y_i, \widehat{Y}_i) \tag{4.1}$$

where L_{CE} is the categorical cross-entropy loss. The first cross-entropy loss compares $\widehat{Y}$, the prediction of the network, and Y, the superpixel mask. The remaining terms compare an intermediate layer prediction $\widehat{Y}_i$ and a downsampled superpixel mask, Y_i (downsampled by a factor of 2^i). The weights of the different intermediate layer losses are controlled by λ_i's.

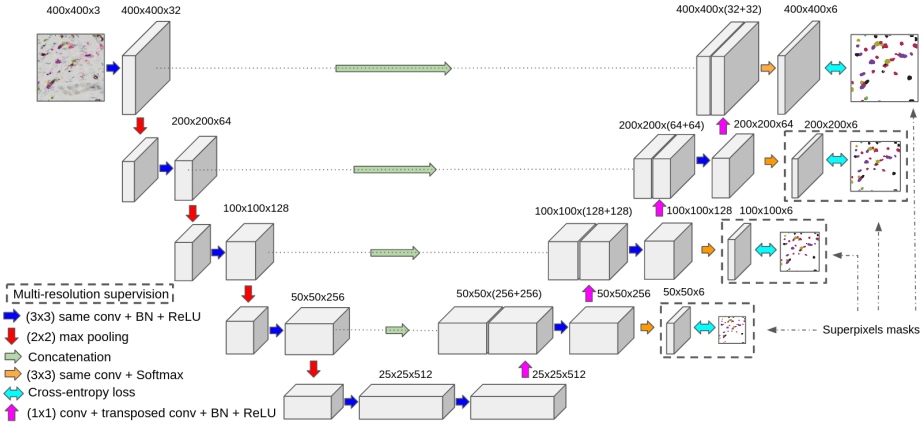


Figure 16: Multi-resolution UNet (MS-UNet).

4.2.2 Ensemble with Color Decomposition Network

The segmentation (or segmentor) network is unable to capture fine cell details, especially near their boundaries. This is inherently unavoidable because the training labels (superpixel masks) are extended automatically from point annotations and are not guaranteed to be accurate. Failing to detect fine scale cell boundaries can be detrimental to downstream analyses that require accurate measurements of pairwise distances between cells, cell sizes and cell distributions.

We observe that the cell boundary is well defined by color chromogenetic stains, which bind to proteins and express themselves in the cell cytoplasm. If we can identify the color stains correctly, we can refine the prediction cell mask. Indeed, color deconvolution (or color decomposition), which is the process of finding masks corresponding to different stain colors, is a classic image analysis task. Various existing color deconvolution methods have been developed [8, 9, 33, 51, 53, 55]. Most of these methods can only solve problems with up to 4 stains. In this paper, we adopt a recent deep autoencoder method that can unmix a large number of stains [1][1].

4.3 Experimental Results

4.3.1 Dataset and Settings.

We evaluated the proposed method using multiplex IHC whole slide images (WSIs) of pancreatic cancer tissue that are stained with chromogenic biomarkers. The biomarkers are CD3 (yellow), CD4 (teal),

[1]The code was obtained by communication with the authors

CD8 (purple), CD16 (black), and CD20 (red), representing different types of immune cells. Fig. 14 depicts instances of these stained immune cell types. The training set consists of 300 patches of size 400×400 pixels randomly sampled from the tumor area of 5 WSIs and the validation set is 60 patches of the same dimensions from another WSI. The test set is 19 patches of size 1200×1920 pixels from 4 different WSIs. The physical resolution of all the patches is 0.174 microns per pixel. The MS-UNet and CDNet are each trained independently on the same dataset. The MS-UNet is trained with the Adam optimizer for 1200 epochs, with an initial learning rate of 0.001 and decreased by 10 every 400 epochs. Using the validation set, the model with the lowest cross entropy loss in only the last output layer is selected. In the final segmentation small connected components that are far less than the typical size of immune cells are removed. A size threshold of 100 pixels, which is around 3 square microns, is empirically chosen.

4.3.2 Quantitative Results.

In our evaluation, F-scores are computed using the point annotations as follows: (i) a true positive is a point that intersects with the prediction segmentation mask of the same label, (ii) a false positive is a connected component in the prediction that does not intersect with a ground truth point of the same label, and (iii) a false negative is a point that does not intersect with any component of the same label in the prediction segmentation mask. We compare the results from the segmentation using CDNet alone, a typical UNet without deep supervision, and MS-UNet with different weights for the deep supervision layers. Performance is improved by varying the weight for the first deep supervision layer (λ_1). Varying the weights for the rest of the remaining deep supervision layers does not improve performance. In Table 1 we compare with different configurations for λ_1, further comparisons with varying λ_2 and λ_3 are in the supplementary material. Table 2 shows that the ensemble *CDNet anchor MS-UNet* method achieves higher performance than the segmentation network UNet and the color decomposition network CDNet. The mean F-score improves by 11.8% and 13% compared to CDNet and UNet respectively. Class-wise, there is a significant improvement in most classes compared to the individual model results in Table 1.

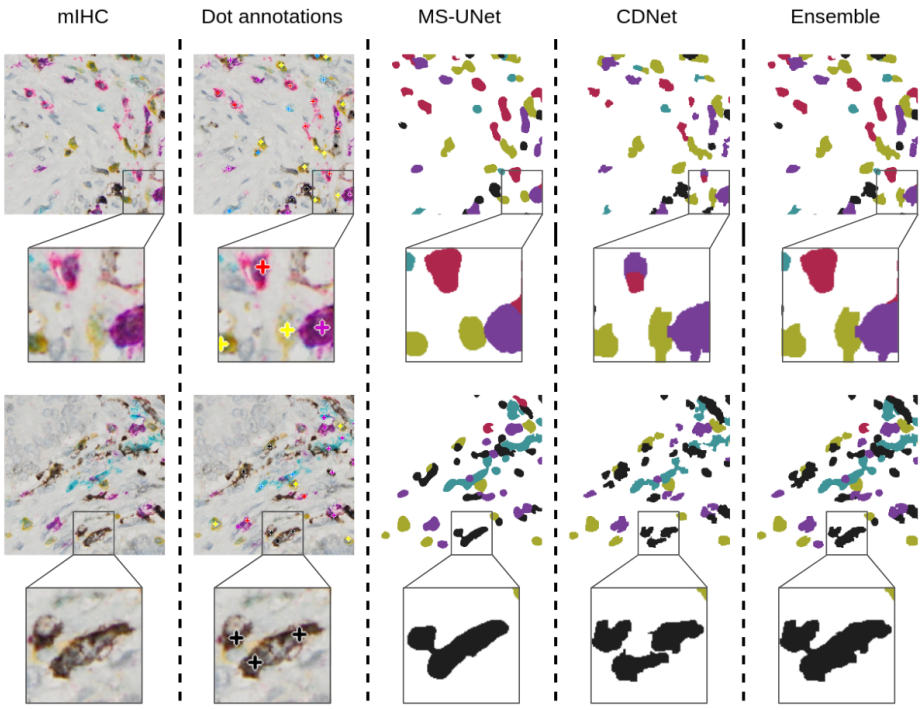


Figure 17: Qualitative results on patches from the test set. Rows 1 and 3 are patches of 400x400 pixels. Rows 2 and 4 are magnifications of subregions of 100x100 pixels.

4.3.3 Qualitative results.

Fig. 17 shows example qualitative results. For each patch, a magnified area is displayed in the following row. In the first magnified sample the red staining in the CD20 cell appears like purple staining. CDNet, which is sensitive to the staining, mistakes part of that cell for CD8, whereas MS-UNet correctly segments the whole cell. In the same magnified region, we see that the yellow staining on the right has a very light color that CDNet picks up more accurately and gives a better segmentation of the whole region than MS-UNet. In the last column we see the advantage of the proposed ensemble *CDNet anchor MS-UNet* which is able to capture the best of both models. In the second magnified sample, MS-UNet mistakenly connects the top cell to the bottom ones. This is due to being trained on coarse superpixels, on the other hand CDNet can better separate the cells. Because we take the union of the 2 results, we leave this issue to future work. Additionally, Fig. 18 shows segmentation predictions on four patches of the test set in full size.

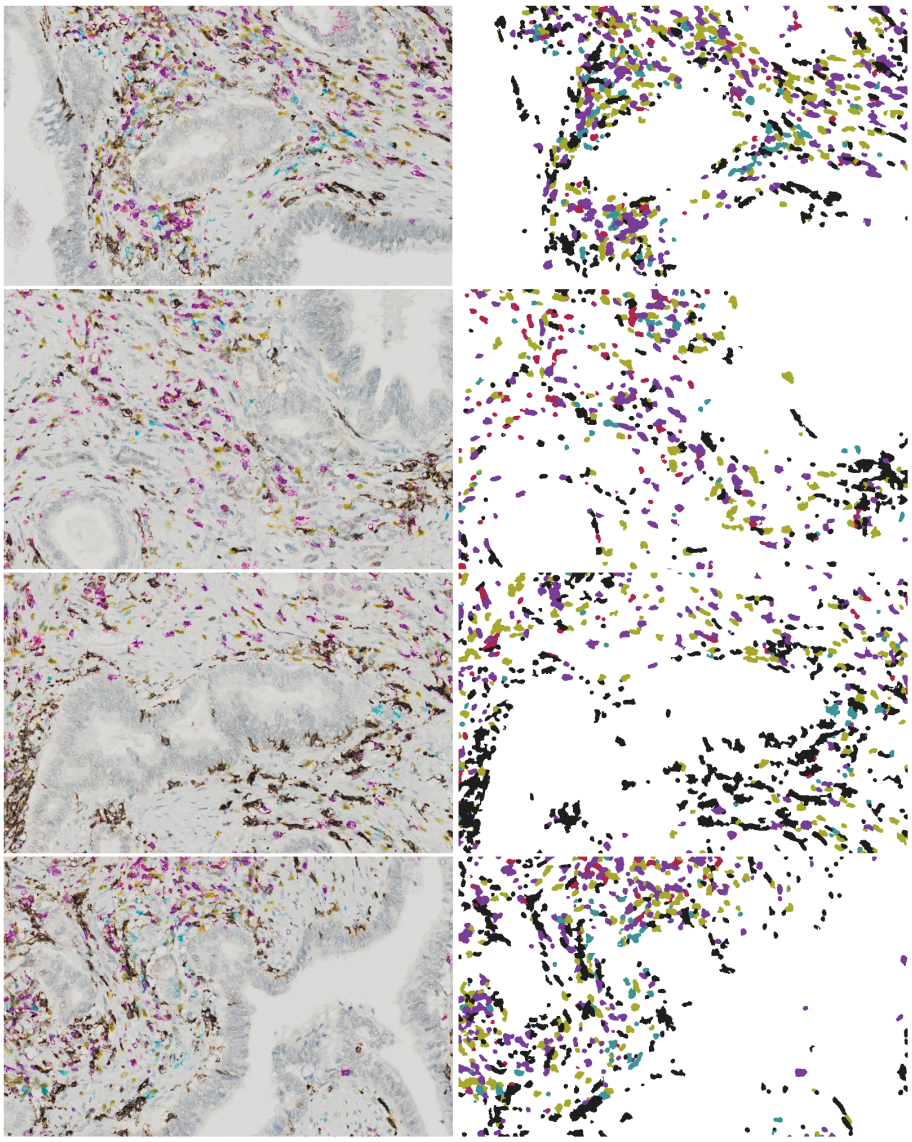


Figure 18: Left: Four full size patches of 1200×1920 pixels from the test set. Right: Segmentation predictions using the ensemble method *CDNet anchor MS-UNet*.

Table 1: Evaluation of color decomposition network and individual segmentation networks

	F1-score					
Method	CD16	CD3	CD4	CD8	CD20	Mean
CDNet	0.6716	0.6168	0.6141	0.6336	0.2166	0.5505
UNet (λ_1=0)	0.7413	0.5876	0.5048	0.6166	0.2738	0.5448
MS-UNet(λ_1=0.50)	0.7383	0.6344	0.5907	0.6248	0.3342	0.5845
MS-UNet(λ_1=0.75)	**0.7473**	**0.6380**	0.5859	**0.6478**	0.3455	**0.5929**
MS-UNet (λ_1=1.00)	0.6881	0.6042	**0.6288**	0.6439	**0.3562**	0.5842

Table 2: Evaluation of ensemble models: color decomposition network (CDNet) and segmentation network (MS-Unet)

	F1-score					
Method	CD16	CD3	CD4	CD8	CD20	Mean
MS-Unet anchor CDNet	0.6775	0.6246	**0.6211**	0.6429	0.2166	0.5565
CDNet anchor MS-UNet	**0.7877**	**0.6651**	0.6079	**0.6618**	**0.3556**	**0.6156**

5 Conclusion and Future Work

We proposed a weakly supervised cell segmentation method for mIHC images. Our method leverages the fine details offered by the color decomposition network and the multi-scale representation learnt through deep supervision. It achieves high quality immune cell segmentation. We expect it to be very useful in downstream quantitative large-scale studies of tumor microenvironments.

The field of weakly supervised learning for semantic segmentation has been developing rapidly over the last years. Lots of works are working with class-level labels and they show impressive results. However, it seems to be that they are starting to saturate on the popular datasets such Pascal VOC 2012. Nowadays, the majority of works do not truly use current large-scale datasets and this can be an opportunity to further reduce the gap between weakly-supervised methods and fully-supervised methods. Also, there is a lack of end-to-end learning methods with most of the weak labels.

From a general view of weakly-supervised problems, when developing a new application, there is a trade off between higher quality annotations and annotation time. Seems to be that recent works are trying to push towards better annotations tools or combining different modalities.

References

[1] S. Abousamra, D. Fassler, L. Hou, Y. Zhang, R. Gupta, T. Kurc, L. F. Escobar-Hoyos, D. Samaras, B. Knudson, K. Shroyer, et al. Weakly-supervised deep stain decomposition for multiplex ihc images. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 481–485. IEEE, 2020.

[2] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012.

[3] A. Balajee Vasudevan, D. Dai, and L. Van Gool. Object referring in videos with language and human gaze. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4129–4138, 2018.

[4] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei. What's the point: Semantic segmentation with point supervision. In *European conference on computer vision*, pages 549–565. Springer, 2016.

[5] S. Blom, L. Paavolainen, D. Bychkov, R. Turkki, P. Mäki-Teeri, A. Hemmes, K. Välimäki, J. Lundin, O. Kallioniemi, and T. Pellinen. Systems pathology by multiplexed immunohistochemistry and whole-slide digital image analysis. *Scientific Reports*, 7(1):1–13, 2017.

[6] Y. B. Can, K. Chaitanya, B. Mustafa, L. M. Koch, E. Konukoglu, and C. F. Baumgartner. Learning to segment medical images with scribble-supervision alone. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 236–244. Springer, 2018.

[7] A. Chaudhry, P. K. Dokania, and P. H. Torr. Discovering class-specific pixels for weakly-supervised semantic segmentation. *arXiv preprint arXiv:1707.05821*, 2017.

[8] T. Chen and C. Chefd'Hotel. Deep learning based automatic immune cell detection for immunohistochemistry images. In *International workshop on machine learning in medical imaging*, pages 17–24. Springer, 2014.

[9] T. Chen and C. Srinivas. Group sparsity model for stain unmixing in brightfield multiplex immunohistochemistry images. *Computerized Medical Imaging and Graphics*, 46:30–39, 2015.

[10] Y.-C. Chen, Y.-Y. Lin, M.-H. Yang, and J.-B. Huang. Show, match and segment: Joint weakly supervised learning of semantic matching and object co-segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[11] J. Dai, K. He, and J. 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.

[12] A. R. Dixon, C. Bathany, M. Tsuei, J. White, K. F. Barald, and S. Takayama. Recent developments in multiplexing techniques for immunohistochemistry. *Expert review of molecular diagnostics*, 15(9):1171–1186, 2015.

[13] Q. Dou, L. Yu, H. Chen, Y. Jin, X. Yang, J. Qin, and P.-A. Heng. 3d deeply supervised network for automated segmentation of volumetric medical images. *Medical image analysis*, 41:40–54, 2017.

[14] Y. Gao, C. Liu, and L. Zhao. Multi-resolution path cnn with deep supervision for intervertebral disc localization and segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 309–317. Springer, 2019.

[15] W. Ge, S. Yang, and Y. Yu. Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1277–1286, 2018.

[16] D. Harwath, A. Recasens, D. Surís, G. Chuang, A. Torralba, and J. Glass. Jointly discovering visual objects and spoken words from raw sensory input. In *Proceedings of the European conference on computer vision (ECCV)*, pages 649–665, 2018.

[17] P. Hofman, C. Badoual, F. Henderson, L. Berland, M. Hamila, E. Long-Mira, S. Lassalle, H. Roussel, V. Hofman, E. Tartour, et al. Multiplexed immunohistochemistry for molecular and immune profiling in lung cancer—just about ready for prime-time? *Cancers*, 11(3):283, 2019.

[18] S. Hong, D. Yeo, S. Kwak, H. Lee, and B. Han. Weakly supervised semantic segmentation using web-crawled videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7322–7330, 2017.

[19] Q. Hou, P. Jiang, Y. Wei, and M.-M. Cheng. Self-erasing network for integral object attention. In *Advances in Neural Information Processing Systems*, pages 549–559, 2018.

[20] R. Hu, M. Rohrbach, and T. Darrell. Segmentation from natural language expressions. In *European Conference on Computer Vision*, pages 108–124. Springer, 2016.

[21] Z. Huang, X. Wang, J. Wang, W. Liu, and J. 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.

[22] B. Jin, M. V. Ortiz Segovia, and S. Susstrunk. Webly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3626–3635, 2017.

[23] S. Karthikeyan, V. Jagadeesh, R. Shenoy, M. Ecksteinz, and B. Manjunath. From where and how to what we see. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 625–632, 2013.

[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.

[25] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár. Panoptic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9404–9413, 2019.

[26] A. Kolesnikov and C. H. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European conference on computer vision*, pages 695–711. Springer, 2016.

[27] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117, 2011.

[28] I. H. Laradji, N. Rostamzadeh, P. O. Pinheiro, D. Vazquez, and M. Schmidt. Where are the blobs: Counting by localization with point supervision. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 547–562, 2018.

[29] I. H. Laradji, N. Rostamzadeh, P. O. Pinheiro, D. Vázquez, and M. Schmidt. Instance segmentation with point supervision. *arXiv preprint arXiv:1906.06392*, 2019.

[30] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. Deeply-supervised nets. In *Artificial intelligence and statistics*, pages 562–570, 2015.

[31] D. Lin, J. Dai, J. Jia, K. He, and J. 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.

[32] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[33] M. Macenko, M. Niethammer, J. S. Marron, D. Borland, J. T. Woosley, X. Guan, C. Schmitt, and N. E. Thomas. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1107–1110. IEEE, 2009.

[34] K.-K. Maninis, S. Caelles, J. Pont-Tuset, and L. Van Gool. Deep extreme cut: From extreme points to object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 616–625, 2018.

[35] S. J. Oh, R. Benenson, A. Khoreva, Z. Akata, M. Fritz, and B. Schiele. Exploiting saliency for object segmentation from image level labels. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pages 5038–5047. IEEE, 2017.

[36] D. P. Papadopoulos, A. D. Clarke, F. Keller, and V. Ferrari. Training object class detectors from eye tracking data. In *European conference on computer vision*, pages 361–376. Springer, 2014.

[37] D. P. Papadopoulos, J. R. Uijlings, F. Keller, and V. Ferrari. Extreme clicking for efficient object annotation. In *Proceedings*

of the IEEE international conference on computer vision, pages 4930–4939, 2017.

[38] D. P. Papadopoulos, J. R. Uijlings, F. Keller, and V. Ferrari. Training object class detectors with click supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6374–6383, 2017.

[39] D. Pathak, P. Krahenbuhl, and T. Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1796–1804, 2015.

[40] D. Pathak, E. Shelhamer, J. Long, and T. Darrell. Fully convolutional multi-class multiple instance learning. *arXiv preprint arXiv:1412.7144*, 2014.

[41] P. O. Pinheiro and R. Collobert. From image-level to pixel-level labeling with convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1713–1721, 2015.

[42] J. Pont-Tuset, J. Uijlings, S. Changpinyo, R. Soricut, and V. Ferrari. Connecting vision and language with localized narratives. In *European Conference on Computer Vision*. Springer, 2020.

[43] X. Qi, Z. Liu, J. Shi, H. Zhao, and J. Jia. Augmented feedback in semantic segmentation under image level supervision. In *European conference on computer vision*, pages 90–105. Springer, 2016.

[44] R. Qian, Y. Wei, H. Shi, J. Li, J. Liu, and T. Huang. Weakly supervised scene parsing with point-based distance metric learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8843–8850, 2019.

[45] H. Qu, P. Wu, Q. Huang, J. Yi, Z. Yan, K. Li, G. M. Riedlinger, S. De, S. Zhang, and D. N. Metaxas. Weakly supervised deep nuclei segmentation using partial points annotation in histopathology images. *IEEE Transactions on Medical Imaging*, 2020.

[46] T. Remez, J. Huang, and M. Brown. Learning to segment via cut-and-paste. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 37–52, 2018.

[47] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pages 817–834. Springer, 2016.

[48] 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.

[49] C. Rother, V. Kolmogorov, and A. Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3):309–314, 2004.

[50] C. Rother, T. Minka, A. Blake, and V. Kolmogorov. Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mrfs. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 993–1000. IEEE, 2006.

[51] A. C. Ruifrok, D. A. Johnston, et al. Quantification of histochemical staining by color deconvolution. *Analytical and quantitative cytology and histology*, 23(4):291–299, 2001.

[52] M. Tang, A. Djelouah, F. Perazzi, Y. Boykov, and C. 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.

[53] D. S. Thommen, V. H. Koelzer, P. Herzig, A. Roller, M. Trefny, S. Dimeloe, A. Kiialainen, J. Hanhart, C. Schill, C. Hess, et al. A transcriptionally and functionally distinct pd-1+ cd8+ t cell pool with predictive potential in non-small-cell lung cancer treated with pd-1 blockade. *Nature medicine*, 24(7):994–1004, 2018.

[54] P. Tokmakov, K. Alahari, and C. Schmid. Weakly-supervised semantic segmentation using motion cues. In *European Conference on Computer Vision*, pages 388–404. Springer, 2016.

[55] A. Vahadane, T. Peng, A. Sethi, S. Albarqouni, L. Wang, M. Baust, K. Steiger, A. M. Schlitter, I. Esposito, and N. Navab. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE transactions on medical imaging*, 35(8):1962–1971, 2016.

[56] A. B. Vasudevan, D. Dai, and L. Van Gool. Object referring in visual scene with spoken language. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1861–1870. IEEE, 2018.

[57] B. Wang, G. Qi, S. Tang, T. Zhang, Y. Wei, L. Li, and Y. Zhang. Boundary perception guidance: A scribble-supervised semantic segmentation approach. In *IJCAI*, pages 3663–3669, 2019.

[58] T. Wang, B. Han, and J. Collomosse. Touchcut: Fast image and video segmentation using single-touch interaction. *Computer Vision and Image Understanding*, 120:14–30, 2014.

[59] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1568–1576, 2017.

[60] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, J. Feng, Y. Zhao, and S. Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2314–2320, 2016.

[61] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, and T. S. Huang. Revisiting dilated convolution: A simple approach for weakly- and semi-supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7268–7277, 2018.

[62] J. Xu, A. G. Schwing, and R. Urtasun. Learning to segment under various forms of weak supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3781–3790, 2015.

[63] K. Yun, Y. Peng, D. Samaras, G. J. Zelinsky, and T. L. Berg. Studying relationships between human gaze, description, and computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 739–746, 2013.

[64] G. Zeng and G. Zheng. Multi-stream 3d fcn with multi-scale deep supervision for multi-modality isointense infant brain mr

image segmentation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 136–140. IEEE, 2018.

[65] Y. Zeng, Y. Zhuge, H. Lu, and L. Zhang. Joint learning of saliency detection and weakly supervised semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7223–7233, 2019.

[66] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[67] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.

[68] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019.

[69] Q. Zhu, B. Du, B. Turkbey, P. L. Choyke, and P. Yan. Deeply-supervised cnn for prostate segmentation. In *2017 International Joint Conference on Neural Networks (Ijcnn)*, pages 178–184. IEEE, 2017.