Noname manuscript No.
(will be inserted by the editor)

A Comprehensive Analysis of Weakly-Supervised Semantic Segmentation in Different Image Domains

Lyndon Chan · Mahdi S. Hosseini · Konstantinos N. Plataniotis

Received: date / Accepted: date

Abstract Recently proposed methods for weakly-supervised semantic segmentation have achieved impressive performance in predicting pixel classes despite being trained with only image labels which lack positional information. Because image annotations are cheaper and quicker to generate, weak supervision is more practical than full supervision for training segmentation algorithms. These methods have been predominantly developed to solve the background separation and partial segmentation problems presented by natural scene images and it is unclear whether they can be simply transferred to other domains with different characteristics, such as histopathology and satellite images, and still perform well. This paper evaluates state-of-the-art weakly-supervised semantic segmentation methods on natural scene, histopathology, and satellite image datasets and analyzes how to determine which method is most suitable for a given dataset. Our experiments indicate that histopathology and satellite images present a different set of problems for weakly-supervised semantic segmentation than natural scene images, such as ambiguous boundaries and class co-occurrence. Methods perform well for datasets they were developed on, but tend to perform poorly on other datasets. We present some practical techniques for these methods on unseen datasets and argue that more work is needed for a generalizable approach to weakly-supervised semantic segmentation. Our full code implementation is available on GitHub: https://github.com/lyndonchan/wsss-analysis.

Keywords Weakly-Supervised Semantic Segmentation · Self-Supervised Learning

Lyndon Chan (✉) · Mahdi S. Hosseini (✉) · Konstantinos N. Plataniotis
The Edward S. Rogers Sr. Department of Electrical & Computer Engineering, University of Toronto, Toronto, Canada
E-mail: lyndon.chan@mail.utoronto.ca
E-mail: mahdi.hosseini@mail.utoronto.ca

1 Introduction

Multi-class semantic segmentation aims to predict a discrete semantic class for every pixel in an image. This is useful as an attention mechanism: by ignoring the irrelevant parts of the image, only relevant parts are retained for further analysis, such as faces and human parts (Prince, 2012a). Semantic segmentation is also useful for changing the pixels of the image into higher-level representations that are more meaningful for further analysis, such as object locations, shapes, sizes, textures, poses, or actions (Shapiro and Stockman, 2000). Oftentimes, semantic segmentation is used when simply predicting a bounding box around the objects is too coarse for fine-grained tasks, especially when the scene is cluttered and the bounding boxes would overlap significantly or when the precise entity boundaries are important. Whereas humans can ordinarily perform such visual inspection tasks accurately but slowly, computers have the potential to perform the same tasks at larger scale and with greater accuracy (Prince, 2012b). Natural scene images can be segmented to monitor traffic density (Audebert et al., 2017), segment humans from images (Xia et al., 2017), and gather crowd statistics (Zhang et al., 2015a). Histopathology images can be segmented to detect abnormally-shaped renal tissues (Kothari et al., 2013), quantify cell size and density (Lenz et al., 2016), and build tissue-based image retrieval systems (Zhang et al., 2015b). Finally, satellite images can be segmented to detect weeds in farmland (Gao et al., 2018), detect flooded areas (Rahnemoonfar et al., 2018), and quantify urban development (Zhang et al., 2019).

The most popular approach to training semantic segmentation models is currently full supervision, whereby the ground-truth pixel segmentation map is observable for training. Fully-supervised semantic segmentation (FSS) methods include OCNet (Yuan et al., 2019), DANet (Fu et al., 2019), HRNet (Wang et al., 2019a), FCN (Long et al., 2015), U-Net

(Ronneberger et al., 2015), sliding window DNN (Ciresan et al., 2012), and multiscale convnet Farabet et al. (2012). Although fully-supervised methods attain state-of-the-art performance, annotating each training image by pixel is costly and slow. The labellers of MS COCO took on average 4.1 seconds to label each image by category and 10.1 minutes to label each image by pixel-level instances (Lin et al., 2014), requiring 150 times the time needed for image-level annotations. Apart from full supervision, other approaches have been proposed to reduce the annotation cost: the unsupervised approach uses unlabelled images, the semi-supervised approach uses a combination of labelled and unlabelled images (or of reliable and noisily-labelled images), and the weakly-supervised approach uses less spatially-informative annotations than the pixel level. Of these approaches, weak supervision generally performs best; its training annotations (in order of decreasing informativeness) include: bounding box (Dai et al., 2015; Papandreou et al., 2015), scribble (Lin et al., 2016), point (Bearman et al., 2016), and image label (Papandreou et al., 2015; Xu et al., 2015). Due to the complete absence of positional information, image-level labels are the cheapest to provide and the most challenging to use, hence this paper will focus on weakly-supervised semantic segmentation (WSSS) from image-level labels.

Numerous fully-supervised methods have already been proposed and have been reported to perform with impressive accuracy. WSSS researchers consider fully-supervised methods to be the "upper-bound" in performance because they are trained with theoretically the most informative supervisory data possible (assuming the annotations are reasonably numerous and accurate) (Kwak et al., 2017; Ye et al., 2018; Kervadec et al., 2019). Indeed, at the time of writing this paper, the best fully-supervised method, DeepLabv3+ (Chen et al., 2018) attained a 89.0% mIoU on the PASCAL VOC2012 test set (Everingham et al., 2010), which is far higher than the current best weakly-supervised method, IRNet (Ahn et al., 2019), with a 64.8% mIoU. Nonetheless, the quality of WSSS methods is impressive, especially considering that learning to segment without any location-specific supervision is an incredibly difficult task - object extents must be inferred solely from their presence in the training images. Qualitatively, existing WSSS methods deliver excellent segmentation performance on natural scene images while requiring only a fraction of the annotation effort needed for FSS. However, since image labels completely lack positional information, weakly-supervised approaches for natural scene images struggle with three major challenges.

Firstly, WSSS methods struggle to differentiate foreground objects from the background, especially if the background contains strongly co-occurring objects, such as the water from *boat* objects, due to the lack of training information on the precise boundary between them. This was observed by (Kolesnikov and Lampert, 2016b; Huang et al., 2018; Zhou et al., 2018) in their qualitative evaluations; (Kolesnikov and Lampert, 2016a) addressed the problem by introducing additional model-specific micro-annotations for training. Secondly, WSSS methods can struggle to differentiate frequently co-occurring foreground objects, such as *diningtable* objects from *chair* objects, especially when the scene is cluttered with overlapping objects or the objects consist of components with different appearance; this was observed by (Kolesnikov and Lampert, 2016b; Zhou et al., 2018). A final challenge is segmenting entire objects instead of discriminative parts, such as the face of a *person* (Zhou et al., 2018). Since CNNs tend to identify only discriminative regions for classification, they only generate weak localization cues at those discriminative parts. Using a CNN with a larger field-of-view has been used to alleviate the problem (Kolesnikov and Lampert, 2016b), while others use adversarial erasing (Wei et al., 2017) or spatial dropout (Lee et al., 2019) to encourage the CNN to identify less-discriminative regions; still others propagate the localization cues out of discriminative parts using semantic pixel affinities (Huang et al., 2018; Ahn et al., 2019).

Furthermore, WSSS methods are typically developed solely for natural scene image benchmark datasets, such as PASCAL VOC2012 and little research exists into applying them to other image domains, apart from (Yao et al., 2016; Nivaggioli and Randrianarivo, 2019) in satellite images and (Xu et al., 2014; Jia et al., 2017) in histopathology images. One might expect WSSS methods to perform similarly after re-training, but these images have many key differences from natural scene images. Natural scene images contain more coarse-grained visual information (i.e. low intra-class variation and high inter-class variation) while satellite and histopathology images contain finer-grained objects (i.e. high intra-class variation and high inter-class variation) (Xie et al., 2019). Furthermore, boundaries between objects are often ambiguous and even experts lack consensus when labelling histopathology (Xu et al., 2017) and satellite images (Mnih and Hinton, 2010), unlike in natural scene images. On the other hand, histopathology and satellite images are always imaged at the same scale and viewpoint with minimal occlusion and lighting variations. These differences suggest that WSSS methods cannot be blindly reapplied to different image domains; it is even possible that an entirely different approach to WSSS might perform better in other image domains.

Previously, we proposed a novel WSSS method called HistoSegNet (Chan et al., 2019), which trains a CNN, extracts weak localization maps, and applies simple modifications to produce accurate segmentation maps on histopathology images. By contrast, WSSS methods developed for natural scene images take the self-supervised learning approach of thresholding the weak localization maps and using them to train a fully-convolutional network. We utilized this approach

because the weak localization maps already corresponded well to the entire ground-truth segments in histopathology images, whereas the authors of other WSSS methods attempted self-supervised learning when they observed their weak localization maps corresponding only to discriminative parts in natural scene images. In this paper, we seek to address the lack of research by applying WSSS to different image domains, especially those which are different from natural scene images and share characteristics with histopathology images. This assessment is crucial to determining whether WSSS can be feasibly applied to certain image domains and to discovering the best practices to adopt in difficult image domains. We make the following three main contributions:

1. We present a comprehensive review of the literature in multi-class semantic segmentation datasets and weakly-supervised semantic segmentation methods from image labels. For each dataset, we explain the image composition and the annotated classes; for each method, we explain the challenges they attempt to solve and the novel approach that they take.
2. We implement state-of-the-art WSSS methods developed for natural scene and histopathology images, and then evaluate them on representative natural scene, histopathology, and satellite image datasets. We conduct experiments to compare their quantitative performance and attempt to explain the results by qualitative assessment.
3. We analyze each approach's compatibility with segmenting different image domains in detail and propose general principles for applying WSSS to different image domains. In particular, we assess: (a) the effect of the sparsity of a classification network's cues, (b) when self-supervised learning is beneficial, and (c) how to address high class co-occurrence in the training data.

The work accomplished in this paper is presented as follows. In Section 2, we present a review of the literature in multi-class semantic segmentation datasets and weakly-supervised semantic segmentation methods from image labels. In Section 3, we present the three representative natural scene, histopathology, and satellite image datasets we selected for evaluation; in Section 4, we present the state-of-the-art WSSS methods to be evaluated and the modifications we used to ensure fair comparison. In Section 5, we analyze their performances quantitatively and qualitatively on the selected datasets. In Section 6, we analyze each approach's compatibility with segmenting different image domains in detail and propose general principles for applying WSSS to different image domains. Finally, our conclusions are presented in Section 7.

2 Related Work

2.1 Multi-class Semantic Segmentation Datasets

We review below the most prominent multi-class semantic segmentation datasets in four image domains: (1) Natural Scene, (2) Histopathology, (3) Visible-light Satellite, and (4) Urban Scene. Each dataset is listed in Table 1; we provide the year of publication, the type of "stuff-things" object annotations, the number of labels per image, the number of classes, the total number of images, the number of pixel-level annotated images, the image size, and optical resolution. Further detailed discussion is provided below.

Fig. 1 Natural Scene Images are captured from natural environments using consumer cameras under vastly varying lighting conditions and viewpoints. The segmentation masks tend to be large and few in number, either leaving large portions of the image unannotated (known as "things"-only, usually for older datasets) or densely covering the entire image (known as "stuff and things", usually for newer datasets) (sample image and ground-truth segmentation from PASCAL VOC2012 (Everingham et al., 2010)).

Natural Scene Images. Natural scene images (also known as "in the wild" or "scene parsing" images) are captured by consumer cameras under varying light conditions and angles. This terminology is used to emphasize that the images are not synthetically-generated or shot under controlled conditions, as image datasets tended to be in the early days of computer vision research. Occlusion, motion blur, cluttered scenes, ambiguous edges, and multiple scales can be present in these images. MSRC-21 (Shotton et al., 2006) is one of the earliest large natural scene datasets annotated at the pixel level, consisting of 591 images (sized $\sim 320 \times 240$), each densely annotated with one or more labels selected from 21 object classes (e.g. *building*, *grass*, *tree*), as well as a *void* class. SIFT Flow (Liu et al., 2010) expanded on the number of annotated images and classes; it consists of 2688 images (all sized 256×256), all annotated with 30 foreground classes (and an *unlabeled* class). PASCAL VOC2012 (Everingham et al., 2010) (see Figure 1) expanded on the number of annotated images even further and subsequently became the benchmark for comparing segmentation algorithms; it consists of 17125 images (with maximum dimension set to 500), 10582 of which are densely annotated with one or more labels selected from 20 foreground classes (e.g. *aeroplane*, *bicycle*,

Name	Classes	# lbl/img	# Classes (fg)	# Img	# GT	Image size	Resolution
Natural Scene Image Datasets							
MSRC-21 (Shotton et al., 2006)	S+T	> 1	21+void	591	591	$\sim 320 \times 240$	Variable
SIFT Flow (Liu et al., 2010)	S+T	> 1	30+unlabeled	2688	2688	256×256	Variable
PASCAL VOC 2012 (Everingham et al., 2010)	T	> 1	20+bg	17125	10582	max $= 500$	Variable
PASCAL-Context (Mottaghi et al., 2014)	S+T	> 1	59	19740	10103	max ≤ 500	Variable
COCO 2014 (Lin et al., 2014)	T	> 1	80	328000	123287	max ≤ 640	Variable
ADE20K (Zhou et al., 2017)	S+T	> 1	2693	22210	22210	median 640×480	Variable
COCO-Stuff (Caesar et al., 2018)	S+T	> 1	172	163957	163957	max ≤ 640	Variable
Histopathology Image Datasets							
C-Path (Beck et al., 2011)	S+T	> 1	9+bg	1286	158	2256×1440	$\sim 0.417\mu$m/px
MMMP (H&E) (Riordan et al., 2015)	S+T	> 1	17+bg	102	15	median 2517×2434	0.321μm/px
HMT (Kather et al., 2016)	S+T	1	7+bg	5000	5000	150×150	0.495μm/px
NCT-CRC (Kather et al., 2019)	S+T	1	8+bg	100000	100000	224×224	0.5μm/px
ADP-morph (Hosseini et al., 2019; Chan et al., 2019)	S+T	> 1	28+bg	17668	50	1088×1088	0.25μm/px
ADP-func (Hosseini et al., 2019; Chan et al., 2019)	S+T	> 1	4+bg+other	17668	50	1088×1088	0.25μm/px
Visible-light Satellite Image Datasets							
UC Merced Land Use (Yang and Newsam, 2010)	S+T	1	21	2100	2100	256×256	1 ft/px
DeepGlobe Land Cover (Demir et al., 2018)	S	> 1	6+unknown	1146	803	2448×2448	50 cm/px
EuroSAT Land Use (Helber et al., 2019)	S+T	1	10	27000	27000	64×64	50 cm/px
Urban Scene Image Datasets							
CamVid (Brostow et al., 2008)	S+T	> 1	31+void	701	701	960×720	Fixed
CityScapes (Cordts et al., 2016)	S+T	> 1	30	5000	3475	2048×1024	Fixed
Mapillary Vistas (Neuhold et al., 2017)	S+T	> 1	66	25000	25000	$\geq 1920 \times 1080$	Fixed
BDD100K (Yu et al., 2018)	S+T	> 1	40+void	100000	10000	1280×720	Fixed
ApolloScape (Wang et al., 2019b)	S+T	> 1	25+unlabeled	146997	146997	3384×2710	Fixed

Table 1 Multi-Class Semantic Segmentation Datasets, listed in chronological order by image domain. "Year" is the year of dataset publication. "Classes" is the type of labelled objects under the "stuff-things" class distinction (T=Things, S=Stuff, S+T=Stuff and Things). "# lbl/img" is the number of labels per image. "# Classes (fg)" is the total number of possible foreground classes. "# Img" is the total number of original images. "# GT" is the number of images provided with pixel-level annotations. "Image size" is the size of the provided original images. "Resolution" is the optical resolution of the camera used to capture the original images.

bird), as well as a *background* class. The original release provided only 1464 pixel-level annotated set called *train*, but these are typically used with an augmented set to form the 10582 pixel-level annotated set called *trainaug* (Hariharan et al., 2011).

PASCAL-Context followed up with a dense annotation of the earlier 2010 release of PASCAL VOC, replacing the *background* class with "stuff" classes (e.g. *road*, *building*, *sky*); it consists of 19740 images (with maximum dimension ≤ 500), 10103 of which are labelled with a more manageable subset of 59 labels. COCO 2014 (Lin et al., 2014) provided an even larger dataset of "thing"-annotated images; it consists of 328000 images (with maximum dimension ≤ 640), 123287 of which are labelled with 80 classes (e.g. *person*, *toilet*, *shoe*), as well as the *background* class. COCO-Stuff (like PASCAL-Context) replaced the *background* class in COCO 2014 with "stuff" classes like *grass* and *sky-other*. ADE20K (Zhou et al., 2017) increases the number of classes considered instead of increasing the number of images contained; it consists of 22210 images (median size 640×480), all of which are densely annotated with 2693 classes (e.g. *door*, *table*, *oven*).

Histopathology Images. Histopathology images are bright-field images of histological tissue slides scanned using a whole slide imaging (WSI) scanner. Although the hematoxylin and eosin (H&E) stain is most commonly used, staining protocols and scanner types often differ between institutions. The scanned slides are themselves tissue cross sections of three-dimensional specimens stained and preserved inside a glass cover and imaged at the same viewpoint. There is no occlusion (except for folding artifacts) and the back-

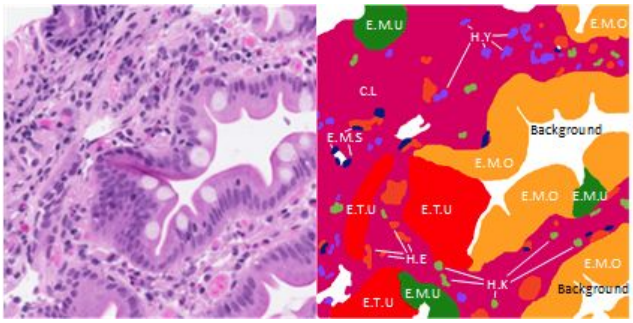


Fig. 2 Histopathology Images are captured from histological tissue slides scanned with a whole slide imaging scanner and are acquired under strictly controlled lighting conditions and viewpoints. The segmentation masks tend to be small and numerous, densely covering the entire image (known as "stuff and things") (sample image and ground-truth segmentation from ADP-morph (Hosseini et al., 2019)).

ground appears uniformly white. Each scanned slide contains vast amounts of visual information, typically to the order of millions of pixels in each dimension. Thus, to reduce the annotation effort, most histopathology datasets are annotated at the patch level rather than the slide level and often each patch is annotated with only one label (Kather et al., 2016, 2019) or with binary classes (Roux et al., 2013; Veta and et.al., 2014; Kumar et al., 2017; Aresta and et.al., 2018). C-Path (Beck et al., 2011) is likely the first histopathology image datasets to annotate at the pixel-level with multiple classes and multiple labels per image; it consists of 1286 patch images (sized 2256×1440), 158 of which are labelled with at least one of 9 histological types (e.g. *epithelial regular nuclei*, *epithelial cytoplasm*, *stromal matrix*) as well as the *background* class. The H&E set of MMMP (Riordan et al., 2015) is smaller, but is annotated with more histological types; it consists of 102 images (median size 2517×2434), 15 of which are annotated with one or more of 17 histological types (e.g. *mitotic figure*, *red blood cells*, *tumor-stroma-nuclear*), as well as the *background* class.

HMT (Kather et al., 2016) and NCT-CRC (Kather et al., 2019) are much larger than C-Path but accomplish this by annotating each image with only one label each. HMT consists of 5000 images (sized 150×150), all labelled with one of 7 histological classes (e.g. *tumour epithelium*, *simple stroma*, *complex stroma*), as well as the *background* class. Ten pixel-level annotated slides (sized 5000×5000) are also provided for evaluation. NCT-CRC consists of 100000 images (sized 224×224), all labelled with one of 8 classes (e.g. *mucus*, *smooth muscle*, *cancer-associated stroma*), as well as the *background* class. ADP (Hosseini et al., 2019; Chan et al., 2019) (see Figure 2) is a histopathology dataset annotated at the pixel level with multiple classes and labels per image; there are 17668 images (sized 1088×1088) in total released with the original dataset (Hosseini et al., 2019). All 17668 images are labelled at the image level, and a subset of 50 images is also annotated as a tuning set in a subsequent paper (Chan et al., 2019) with 28 morphological types (known as "ADP-morph") and 4 functional types (known as "ADP-func"). A different subset of 50 images is annotated as an evaluation set and presented in this paper.

Visible-Light Satellite Images. Visible-light satellite images are images of the Earth taken in the visible-light spectrum by satellites or airplanes. Typically, the surface of the Earth is the object of interest, although occlusion by atmospheric objects (such as clouds) is not uncommon. Lighting conditions can vary, depending on the time of day, and the viewpoint tends not to vary significantly for objects directly below the satellite (distant objects experience distortion due to parallax). Like histopathology images, each satellite image contains vast amounts of visual information, so most satellite image datasets are annotated at the patch level to reduce the annotation cost. UC Merced Land Use (Yang and Newsam, 2010) and EuroSat Land Use (Helber et al., 2019) are both annotated with a single label per image. UC Merced Land Use consists of 2100 images (sized 256×256), each labelled with one of 21 land use classes (e.g. *agricultural*, *denseresidential*, *airplane*). EuroSat Land Use, on the other hand, consists of 27000 images (sized 64×64), each labelled with one of 10 land use classes (e.g. *AnnualCrop*, *Industrial*, *Residential*). DeepGlobe Land Cover (Demir et al., 2018) (see Figure 3) was released for a fully-supervised semantic segmentation challenge and is annotated with multiple labels per image; it comprises of 1146 images (sized 2448×2448), 803 of which are annotated with one or more of 6 classes (e.g. *urban*, *agriculture*, *rangeland*), as well as an *unknown* class.

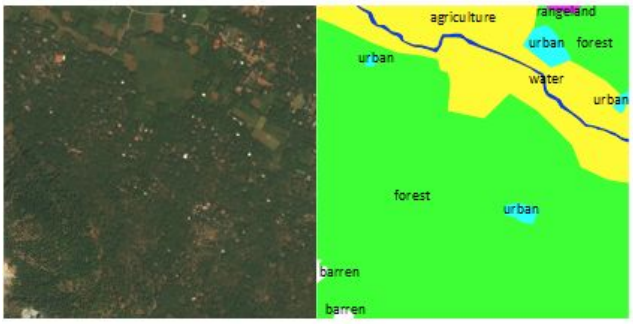


Fig. 3 Visible-Light Satellite Images are captured from the planetary surface using satellites or airplanes under mildly varying lighting conditions and viewpoints. The segmentation masks tend to be small and numerous, densely covering the entire image (known as "stuff and things") (sample image and ground-truth segmentation from DeepGlobe Land Cover (Demir et al., 2018)).

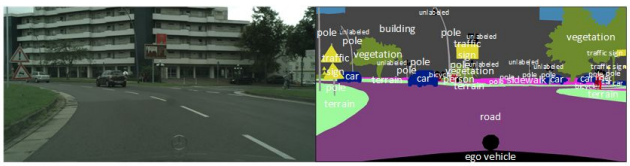


Fig. 4 Urban Scene images are captured from city environments using a car-mounted camera under vastly varying lighting conditions and viewpoints. The segmentation masks tend to be medium-sized and numerous, densely covering the entire image (known as "stuff and things") (sample image and ground-truth segmentation from CityScapes (Cordts et al., 2016)).

Urban Scene Images. Urban scene images are images of scenes in front of a driving car, captured by a fixed surveillance camera mounted behind the windshield. Typically, images are captured under different lighting conditions while street-level viewpoint can vary; occlusion is a possibility. The first major urban scene dataset was CamVid (Brostow et al., 2008), which densely annotated all 701 images (sized 960×720) with one or more than labels of 31 urban scene

classes (e.g. *Bicyclist*, *Building*, *Tree*), as well as *void* class. CityScapes (Cordts et al., 2016) (see Figure 4) consists of 5000 images (sized 2048×1024), 3475 of which are annotated with 30 classes. Mapillary Vistas (Neuhold et al., 2017) is even larger; it consists of 25000 images (sized at least 1920×1080), all annotated with 66 object categories (for semantic segmentation). BDD100K (Yu et al., 2018) consists of a larger set of 100000 images (sized 1280×720), but only 10000 of these are annotated for instance segmentation with 40 object classes (and a *void* class). The April 3, 2018 release of ApolloScape (Wang et al., 2019b) is the largest of all to date; it consists of 146997 images (sized 3384×2710), all annotated at the pixel level with 25 classes (and an *unlabeled* class).

2.2 Weakly-Supervised Semantic Segmentation

Below, we review the literature in weakly-supervised semantic segmentation from image-level annotations, which refers to learning pixel-level segmentation from image-level labels only. This is the least informative form of weak supervision available for semantic segmentation as it provides no location information for the objects. Different WSSS methods trained with image-level annotations have been proposed to solve this problem; their methodologies can be broadly categorized into four approaches: Expectation-Maximization, Multiple Instance Learning, Object Proposal Class Inference, and Self-Supervised Learning. Table 2 organizes the reviewed methods by their approaches and common features, while Table 3 lists the methods chronologically with information on the availability of their code online and their segmentation performance in PASCAL VOC 2012, which most of them were developed for.

(1) Expectation-Maximization. The Expectation - Maximization approach consists of alternately optimizing a latent label distribution across the image and learning a segmentation of the image from that latent distribution. In practice, this means starting with a prior assumption about the class distribution (e.g. the size of each class segment) from the ground-truth image annotations, training a Fully Convolutional Network (FCN) to replicate these inferred segments, updating the prior assumption model based on the FCN features, and repeating the training cycle again.

CCNN (Pathak et al., 2015) uses block coordinate descent to alternate between (1) optimizing the convex latent distribution of fixed FCN outputs with segment-specific constraints (e.g. for suppressing absent labels and encouraging large foreground segments) and (2) training a FCN with SGD against the fixed latent distribution. EM-Adapt (Papandreou et al., 2015) alternates between (1) training a FCN with class-specific bias to each activation map with global sum pooling on the log activation maps to train against the image-level labels and (2) adaptively setting the class biases to equal a fixed percentile of the score difference between the maximum and class score at each position (in order to place a lower bound on the segment area of each class).

(2) Multiple Instance Learning. The Multiple Instance Learning (MIL, or Bag of Words) approach consists of learning to predict the classes present in an image (known as a "bag") given ground-truth image-level annotations and then, given the knowledge that at least one pixel of each class is present, assigning pixels (known as "words") to each predicted class. In practice, this often means training a Convolutional Neural Network (CNN) with image-level loss and inferring the image locations responsible for each class prediction.

MIL-FCN (Pathak et al., 2014) trains a FCN headed by a 1×1 conv layer and a Global Max Pooling (GMP) layer against the image-level annotations, then at test time, it predicts the top class at each location in the convolutional features and the predicted class map is bilinearly upsampled. DCSM (Shimoda and Yanai, 2016) trains a CNN at the image level and uses GBP (guided back-propagation) to obtain the coarse class activation maps at the upper intermediate convolutional layers, then subtracts the maps from each other, and takes the average of the maps across different scales and layers, followed by CRF post-processing. BFBP (Saleh et al., 2016) trains a FCN with a foreground/background mask generated by CRF on the scaled average of conv4 and conv5 features with cross-entropy loss between the image-level annotations and the LSE pool of foreground- and background-masked features; CRF post-processing is applied at test time. WILDCAT (Durand et al., 2017) trains a FCN with conv5 features being fed into a WSL transfer network, then applies class-wise average pooling and weighted spatial average of top- and lowest-activating activations; at test time, it infers the maximum-scoring class per position and post-processes with CRF.

(3) Object Proposal Class Inference. The Object Proposal Class Inference approach often takes elements from both the MIL and Self-Supervised Learning approaches but starts by extracting low-level object proposals and then assigns the most probable class to each one using coarse-resolution class activation maps inferred from the ground-truth image-level annotations. SPN (Kwak et al., 2017) trains a CNN which performs a spatial average of the features closest to each superpixel from the original image and then has FC classifier layers with an image-level loss, and these superpixel-pooled features are then used as pseudo ground-truths to train a FCN. PRM (Zhou et al., 2018) extracts MCG (Multi-scale Combinatorial Grouping) low-level object proposals, trains a FCN with peak stimulation loss, then peak backpropagation is done for each peak in the Class Response Map to obtain the Peak Response Map. Each object proposal is then scored using the PRM peaks and assigned the top-ranked classes with non-maximum suppression.

Method	Fully-supervised classification net	Spatial dropout	Expectation-Maximization	CLM inference	Fine contour modification	CLM propagation	Object proposal class inference	Self-supervised segmentation net	Method description
(1) Expectation-Maximization Methods									
CCNN (Pathak et al., 2015)	•		•	•			•		Optimize convex latent distribution as pseudo GT; train FCN + CRF
EM-Adapt (Papandreou et al., 2015)	•		•	•			•		Train FCN + predict with class-specific bias to log activation maps + CRF
(2) Multiple Instance Learning Methods									
MIL-FCN (Pathak et al., 2014)	•			•					Train FCN w/ GMP + predict with top prediction at each location, upsample
DCSM (Shimoda and Yanai, 2016)	•			•	•				Train CNN + GBP + depth max + class subtract + multi-scale/layer avg + CRF
BFBP (Saleh et al., 2016)	•			•	•				Train CNN w/ avg of conv4/5 + fg/bg mask + CRF + LSE
WILDCAT (Durand et al., 2017)	•			•	•				Train CNN + class avg of conv feature + pool + local predict + CRF
(3) Object Proposal Class Inference Methods									
SPN (Kwak et al., 2017)	•			•			•	•	Train CNN against GAP and SP as pseudo GT + train FCN
PRM (Zhou et al., 2018)	•			•			•		Train CNN w/ PSL + CRM + PB to PRM + predict class for each MCG proposal
(4) Self-Supervised Learning Methods									
SEC (Kolesnikov and Lampert, 2016b)	•			•	•			•	Train CNN + CAM as pseudo GT + train FCN + predict with CRF
MDC (Wei et al., 2018)	•			•				•	Train CNN + avg multi-dilated CAM + weigh w/ scores as pseudo GT + train FCN
AE-PSL (Wei et al., 2017)	•	•		•				•	Erase DOR during CNN training + CAM as pseudo GT + train FCN
FickleNet (Lee et al., 2019)	•	•		•				•	Train CNN w/ dropout in conv RF + repeat Grad-CAM as pseudo GT + train FCN
DSRG (Huang et al., 2018)	•			•	•	•		•	Train CNN + CAM + region growing as pseudo GT + train FCN + predict with CRF
PSA (Ahn and Kwak, 2018)	•			•	•	•		•	Train CNN + CAM + random walk in SAG + CRF as pseudo GT + train FCN
IRNet (Ahn et al., 2019)	•			•	•	•		•	Train CNN + CAM + RW in CAM from centroids as pseudo GT + train FCN

Table 2 Weakly-Supervised Semantic Segmentation Methods, organized by approach, with common methodological features and short description for each method.

Method	Year	Code available?	Train/test code	Code framework	VOC2012 mIoU (%) *val*	VOC2012 mIoU (%) *test*
MIL-FCN (Pathak et al., 2014)	2015	Y	Train/test	MatConvNet	25.7	24.9
CCNN (Pathak et al., 2015)	2015	Y	Train/test	Caffe	35.3	35.6
EM-Adapt (Papandreou et al., 2015)	2015	Y: Caffe, TensorFlow	Train/test	Caffe, TensorFlow	**38.2**	**39.6**
DCSM w/o CRF (Shimoda and Yanai, 2016)	2016	Y	Test	Caffe	40.5	41
DCSM w/ CRF (Shimoda and Yanai, 2016)	2016	Y	Test	Caffe	44.1	45.1
BFBP (Saleh et al., 2016)	2016	N	No	-	46.6	48.0
SEC (Kolesnikov and Lampert, 2016b)	2016	Y: Caffe, TensorFlow	Train/test	Caffe, TensorFlow	**50.7**	**51.7**
WILDCAT + CRF (Durand et al., 2017)	2017	Y	Train/test	PyTorch	43.7	-
SPN (Kwak et al., 2017)	2017	Y	Custom layer only	Keras	50.2	46.9
AE-PSL (Wei et al., 2017)	2017	N	No	-	**55.0**	**55.7**
PRM (Zhou et al., 2018)	2018	Y	Test	PyTorch	53.4	-
DSRG (VGG16) (Huang et al., 2018)	2018	Y: Caffe, TensorFlow	Train/test	Caffe, TensorFlow	59.0	60.4
PSA (DeepLab) (Ahn and Kwak, 2018)	2018	Y	Train/test	PyTorch	58.4	60.5
MDC (Wei et al., 2018)	2018	N	No	-	60.4	60.8
DSRG (ResNet101) (Huang et al., 2018)	2018	Y: Caffe, TensorFlow	Train/test	Caffe, TensorFlow	61.4	63.2
PSA (ResNet38) (Ahn and Kwak, 2018)	2018	Y	Train/test	PyTorch	**61.7**	**63.7**
FickleNet (Lee et al., 2019)	2019	N	No	-	61.2	61.9
IRNet (Ahn et al., 2019)	2019	Y	Train/test	PyTorch	**63.5**	**64.8**

Table 3 Weakly-Supervised Semantic Segmentation Methods (developed for PASCAL VOC2012), by year of publication from 2015 to 2019. Code availability and performance on the PASCAL VOC2012 *val* and *test* sets are also provided for each method.

(4) Self-Supervised Learning. The Self-Supervised Learning approach is similar to the MIL approach but uses the inferred pixel-level activations as pseudo ground-truth cues (or seeds) for self-supervised learning of the final pixel-level segmentation maps. In practice, this usually means training a "backbone" classification network to produce Class Activation Map (CAM) seeds and then training a FCN segmentation network on these seeds. SEC (Kolesnikov and Lampert, 2016b) is the prototypical method to take this approach; it trains a CNN and applies CAM to produce pseudo ground-truth segments to train a FCN against the generated seeds, against the image-level label, and a constraint loss against the CRF-processed maps. MDC (Wei et al., 2018) takes a similar but more multi-scale approach by training a CNN with multi-dilated convolutional layers at the image level, adding multi-dilated block CAMs together, and then generating pseudo ground-truths to train a FCN with the class score-weighted maps. However, methods taking this approach tend to produce good segmentations only for discriminative parts rather than entire objects, so different solutions have been suggested to fill the low-confidence regions in between.

One solution is to apply adversarial or stochastic erasing during training and encourage the networks to learn less discriminative object parts. AE-PSL (Wei et al., 2017) generates CAMs as pseudo ground-truths for training a FCN just like SEC, but during CNN training, high-activation regions from the CAMs are adversarially erased from the training image. FickleNet (Lee et al., 2019), on the other hand, trains a CNN at the image level with centre-fixed spatial dropout in the later convolutional layers (by dropping out non-centre pixels in each convolutional window) and then runs Grad-CAM multiple times to generate a thresholded pseudo ground-truth for training a FCN.

Another solution is to simply propagate class activations from high-confidence regions to adjacent regions with similar visual appearance. DSRG (Huang et al., 2018) trains a CNN and applies region-growing on the generated CAMs to produce a pseudo ground-truth for training a FCN. PSA (Ahn and Kwak, 2018) similarly trains a CNN but propagates the class activations by performing a random walk from the seeds in a semantic affinity graph as a pseudo ground-truth for training a FCN. IRNet (Ahn et al., 2019) is similar as well, but seeks to segment individual instances by performing the random walk from low-displacement field centroids in the CAM seeds up until the class boundaries as the pseudo ground-truths for training a FCN. It is significant to note that, judging from their quantitative performance on PASCAL VOC2012-*val*, the top five performing WSSS methods all use the self-supervised learning approach, and three of these additionally use the outward class propagation technique.

2.3 Semantic Segmentation Methods for Satellite and Histopathology Images

Satellite Images. Compared to natural scene images, relatively limited research has been conducted in multi-class semantic segmentation in satellite images. Most work has been done with fully-supervised learning, since these annotations are the most informative. Indeed, the best performing methods tend to use variants of popular methods developed for natural scene images. In the DeepGlobe Land Cover Classification challenge (Demir et al., 2018), for instance, DFCNet (Tian et al., 2018) is the best performing method and is a variant of the standard FCN (Long et al., 2015) with multi-scale dense fusion blocks and auxiliary training on the road segmentation dataset. The second-best method Deep Aggregation Net (Kuo et al., 2018) is DeepLabv3 (Chen et al., 2017b) with Gaussian filtering applied to the segmentation masks and graph-based post-processing to remove small segments (by assigning them to the class of their top-left neighbouring segment if their size falls below a threshold). The third-best method (Seferbekov et al., 2018) uses a variant of FPN (Lin et al., 2017b), but the convolutional branch networks attached to the intermediate convolutional layers (known as RPN heads in the original FPN method for proposing object regions) with skip connections are instead used to output multi-scale features that are concatenated into a final segmentation map (at the original image resolution). Another assessment of different semantic segmentation techniques on the even larger NAIP dataset used the standard DenseNet and U-Net architectures without significant modifications (Robinson et al., 2019). For weakly-supervised learning, even less research is published; what research can be found attempts to apply standard WSSS techniques to satellite images. Indeed, the state-of-the-art Affinity-Net (or PSA) was adapted by (Nivaggioli and Randrianarivo, 2019) for segmenting DeepGlobe images with only image-level annotations (while experimenting with de-emphasizing background loss and eliminating the background class altogether). SDSAE (Yao et al., 2016)was used to train on image-level land cover annotations on the LULC set as auxiliary data and the trained parameters were then transferred to perform pixel-level segmentation on their proposed Google Earth land cover dataset.

Histopathology Images. In histopathological images, semantic segmentation methods tend to address binary-class problems, probably due to the significant expense of annotating large histopathology images with multiple classes. These tend to label each pixel with either diagnoses (e.g. cancer/non-cancer (Aresta and et.al., 2018)) tissue/cell types (e.g. gland (Sirinukunwattana et al., 2017), nuclei (Kumar et al., 2017), and mitotic/non-mitotic figures (Roux et al., 2013; Veta and et.al., 2014)). As with satellite imagery, semantic segmentation methods for histopathology tend to use

fully-supervised learning. Sliding patch-based methods have been used to segment mitotic figures (Cireşan et al., 2013; Malon and Cosatto, 2013), cells (Shkolyar et al., 2015), neuronal membranes (Ciresan et al., 2012), and glands (Li et al., 2016; Kainz et al., 2015). Superpixel-based object proposal methods have been used to segment tissues by histological type (Xu et al., 2016; Turkki et al., 2016). Fully convolutional methods have been used by training a FCN with optional contour post-processing (Chen et al., 2016; Lin et al., 2017a). Weakly-supervised methods, on the other hand, are much rarer and tend to use a patch-based MIL approach. MCIL was developed to segment colon TMAs by cancer grade with only image-level annotations by clustering the sliding patch features (Xu et al., 2014). EM-CNN (Hou et al., 2016) is trained on slide-level cancer grade annotations and predicts at the patch level and forms a decision fusion model afterward to predict the cancer grade of the overall slide. Although the pre-decision segmentation map only has patch-level resolution, it could theoretically be extended to pixel-level resolution had the patches been extracted densely at test time. DWS-MIL (Jia et al., 2017) trains a binary-class CNN with multi-scale loss against the image-level labels by assuming the same label throughout each ground-truth image (essentially using Global Average Pooling (GAP)). ScanNet (Lin et al., 2018) is a FCN variant trained on patch-level prediction; at test time, a block of multiple patches is inputted to the network and a coarse pixel-level segmentation is outputted; originally developed for breast cancer staging, it has also been applied to lung cancer classification (Wang et al., 2018). HistoSegNet (Chan et al., 2019) trains a CNN on patch-level histological type annotations and applies Grad-CAM to infer coarse class maps, followed by class-specific modifications (*background* and *other* class map augmentation, class map subtraction), and post-processing with CRF to produce fine pixel-level segmentation maps.

3 Datasets

At the time of writing, the vast majority of WSSS algorithms have been developed for natural scene images. Hence, to analyze their performance on other image domains, we selected three representative datasets for evaluation: (1) Atlas of Digital Pathology (histopathology), (2) PASCAL VOC2012 (natural scene), and (3) DeepGlobe Land Cover Classification (satellite).

3.1 Atlas of Digital Pathology (ADP)

The Atlas of Digital Pathology (Hosseini et al., 2019) is a database of histopathology patch images (sized 1088×1088) extracted from WSI scans of healthy tissues stained by the same institution and scanned from different organs with the Huron TissueScope LE1.2 scanner (0.25μm/pixel resolution). This dataset was selected due to the large quantity of image-labelled histopathology patches available for training, each labelled with 28 morphological types (with *background* added for segmentation) and 4 functional types (with *background* and *other* added for segmentation). We use the *train* set of 14,134 image-annotated patches for training; for validation, we use the *tuning* set of 50 pixel-annotated patches, which has more classes per image and was to tune HistoSegNet (Chan et al., 2019); for evaluation, we use the *segtest* set of 50 pixel-annotated patches.

3.2 PASCAL VOC2012

The 2012 release of the PASCAL VOC challenge dataset (Everingham et al., 2010) consists of natural scene ("in the wild") images captured by a variety of consumer cameras. This dataset was selected due to its status as the default benchmark set for WSSS algorithms. Each image is labelled with 20 foreground classes, with an added *background* class for segmentation. For training, we use the *trainaug* set of 12,031 image-annotated images (Hariharan et al., 2011); for evaluation, we use the *val* set of 1,449 pixel-annotated images (the segmentation challenge ranks methods with the *test* set of 1,456 un-annotated images through the evaluation server).

3.3 DeepGlobe Land Cover Classification

The DeepGlobe Land Cover Classification dataset consists of visible-light satellite images extracted from the DigitalGlobe+Vivid Images dataset (Demir et al., 2018). This dataset was selected due to its status as the only multi-label satellite dataset for segmentation. Each image is labelled with 6 land cover classes (and an *unknown* class for non-land cover regions). For training, we randomly split the *train* set of 803 pixel-annotated images into our own 75% training set of 603 image-annotated images and 25% test set of 200 pixel-annotated images. The *unknown* class was omitted for both training and evaluation.

4 Methods

To compare WSSS algorithm performance on the selected datasets, three state-of-the-art methods were chosen: (1) SEC, (2) DSRG, and (3) HistoSegNet. SEC and DSRG were both developed for natural scene images (PASCAL VOC2012) and had both the highest mean Intersection-over-Union (mIoU) at the time of writing and had code implementations available online; HistoSegNet was developed for histopathology images (ADP) and is the only WSSS method developed specifically for non-natural scene images. Furthermore, SEC

and DSRG share a common self-supervised FCN training approach while HistoSegNet uses a simpler Grad-CAM refinement approach. See 5 for an overview of the three evaluated methods.

4.1 Seed, Expand and Constrain (SEC)

Seed, Expand and Constrain (SEC) (Kolesnikov and Lampert, 2016b) was developed for the PASCAL VOC2012 dataset and consists of four trainable stages: (1) a classification CNN is trained on image labels, (2) CAMs are generated from the trained CNN, (3) the CAMs are thresholded and overlap conflicts resolved as seeds/cues, and (4) the seeds are used for self-supervised training of a FCN (DeepLabv1, also known as DeepLab-LargeFOV (Chen et al., 2014)).

(1) Classification CNN. First, two classification CNNs are trained on the annotated images: (1) the "foreground" network (a variant of the VGG16 network omitting the last two pooling layers and the last two fully-connected layers and replacing the flattening layer with a GAP layer) and (2) the "background" network (a variant of the VGG16 network omitting the last two convolutional blocks).

(2) CAM. The Class Activation Map (CAM) is then applied to both the "foreground" and "background" networks for each image in the *trainaug* dataset.

(3) Seed Generation. For the "foreground" network, each class CAM is thresholded above 20% of the maximum activation as a weak localization cue (or seed); for "background" network, the class CAMs are added, a 2D median filter is applied, and the 10% lowest-activating pixels are thresholded as the additional *background* cue. In regions where cues overlap, the class with the smaller cue takes precedence.

(4) Self-Supervised FCN Learning. Finally, these weak localization cues are used as pseudo ground-truths for self-supervised learning of a Fully Convolutional Network (FCN) (Long et al., 2015). A three-part loss function is used on the FCN output: (1) a seeding loss with the weak cues, (2) an expansion loss with the image labels, and (3) a constrain loss with itself after applying dense CRF. At test time, dense CRF is used for post-processing.

4.2 Deep Seeded Region Growing (DSRG)

Deep Seeded Region Growing (DSRG) (Huang et al., 2018) was, similarly to SEC, also developed for PASCAL VOC2012 and takes the similar approach of generating weak seeds using CAM for training a FCN (this time, DeepLabv2, also known as DeepLab-ASPP (Chen et al., 2017a)). However, this method differs in several important ways. First, there is no "background" network - the *background* activation is instead generated separately using the fixed DRFI method (Jiang et al., 2013). Secondly, the foreground CAMs are thresholded above 20% of the maximum activation and then used as seeds for convolutional feature-based region growing into a weak localization cue. Thirdly, a two-part loss function is used on the FCN output: (1) a seeding loss with the region-grown weak cues and (2) a boundary loss with itself after applying dense CRF (identical to constrain loss in SEC). Again, dense CRF is applied at test time.

4.3 IRNet

Inter-pixel Relation Network (IRNet) (Ahn et al., 2019) was developed for both semantic and instance segmentation in PASCAL VOC2012, although we only consider the semantic segmentation case. While it utilizes CAMs as pseudo ground-truths like SEC and DSRG, it trains two branches from the backbone network to predict auxiliary information instead of the pixel classes directly. The method consists of the following five stages:

(1) Classification CNN & (2) CAM. Similarly to SEC and DSRG, a classification CNN is first trained on the labelled images (ResNet50 architecture (He et al., 2016)) and CAMs are generated after training is complete.

(3) Seed Generation. The CAM of each confident class is then thresholded above 0.3 and refined with dense CRF as foreground seeds. Regions with CAM confidence below 0.05 and left without a foreground seed after refining with dense CRF are considered as background seeds.

(4) Self-Supervised DF and CBM Learning. The foreground and background seeds are used as pseudo ground-truths for training two branches from the backbone network: (1) a displacement field (DF) to predict the positional displacement of each pixel from each seed instance's centroid and (2) a class boundary map (CBM) to predict the likelihood of a class boundary existing at each pixel, by maximizing the value between neighbouring pixels (within a set radius) seeded with different classes and minimizing it for pixels of the same seed class.

(5) CAM Random Walk Propagation with CBM. Finally, the CAM of each confident class is propagated by random walk with the inverse of the class boundary map as the transition probability matrix. This enables confident CAM regions to propagate into less confident regions inside likely class boundaries.

4.4 HistoSegNet

The HistoSegNet algorithm (Chan et al., 2019) was developed for the ADP database of histological tissue type (HTT), and consists of four stages: (1) a classification CNN is trained on patch-level annotations, followed by (2) a hand-crafted Grad-CAM, (3) activation map adjustments (e.g. *background*

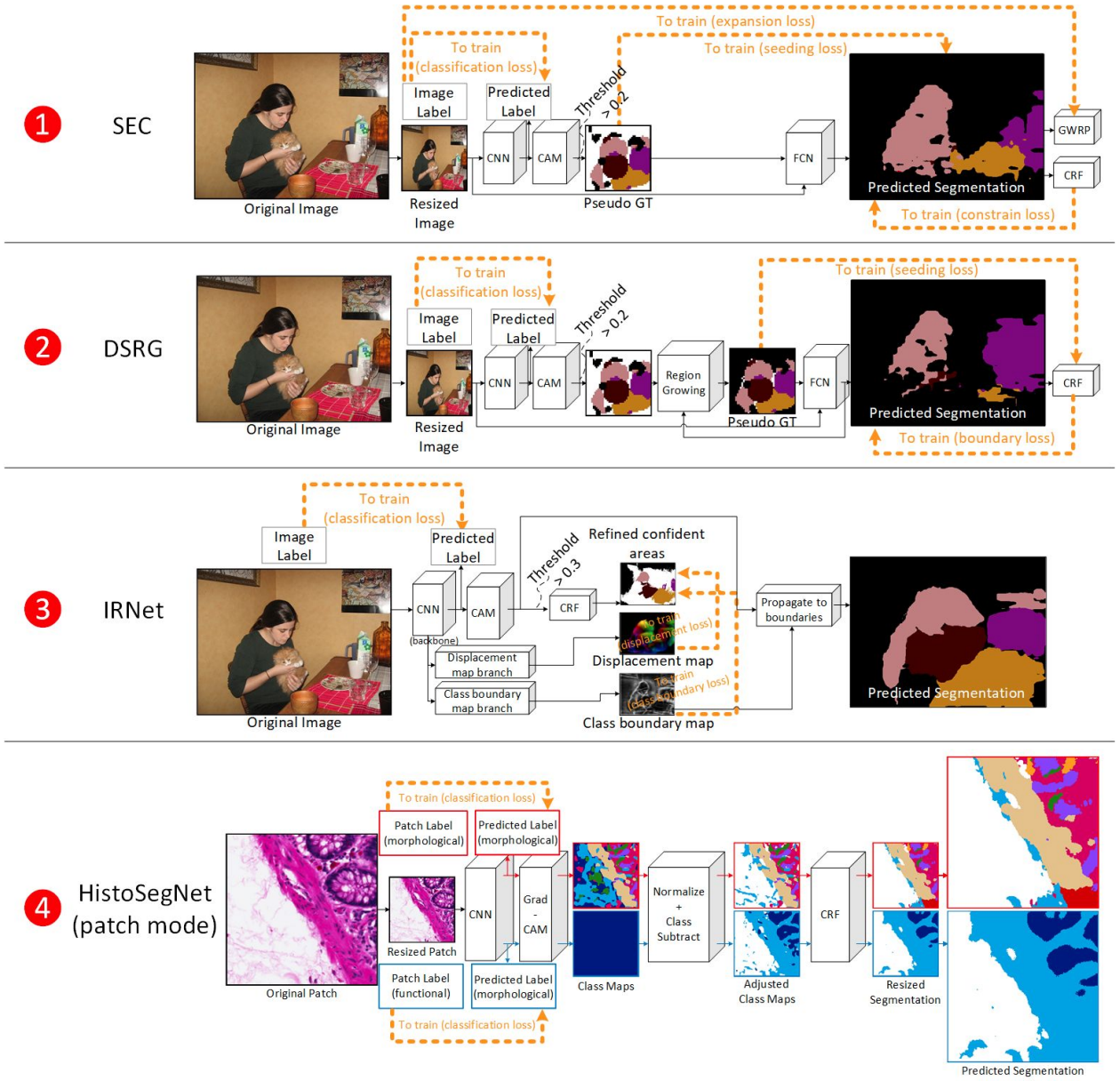


Fig. 5 Overview of the four compared WSSS methods: (1) SEC, (2) DSRG, (3) IRNet, and (4) HistoSegNet. All four methods first train a classification CNN on the image labels, produce coarse activation maps, and then process these maps to infer fine-grained pixel labels. SEC, DSRG, and IRNet are self-supervised learning methods developed for PASCAL VOC2012; they produce fine-grained segmentations by training a downstream neural network model with the coarse maps as pseudo ground-truths. HistoSegNet is an object proposal class inference method developed for ADP; it produces fine-grained segmentation by applying a hand-tuned dense conditional random field to the coarse maps.

/ *other* activations, class subtraction), and (4) a dense CRF. By default, HistoSegNet accepts 224×224-pixel patches that are resized from a scan resolution of $0.25 \times \frac{224}{1088} = 1.2143\mu$m/pixel. Processing is conducted mostly independently for the morphological and functional segmentation modes. Patch predictions between stages (3) and (4) to minimize boundary artifacts.

(1) Classification CNN. First, a classification CNN is trained on the HTT-labelled patches of the ADP database (i.e. the 31 HTTs in the third level, excluding undifferentiated and absent types). The architecture is a variant of VGG-16, except: (1) the softmax layer is replaced by a sigmoid layer, (2) batch normalization is added after each convolutional layer activation, and (3) the flattening layer is replaced by a global max pooling layer. Furthermore, no color normalization was applied since the same WSI scanner and staining protocol were used for all images.

(2) Grad-CAM. To infer pixel-level HTT predictions from the pre-trained CNN, Gradient-Weighted Class Activation Maps (Grad-CAM) (Selvaraju et al., 2017) are applied;

this is a generalization of Class Activation Map (CAM) (Zhou et al., 2016) for all CNN architectures. Grad-CAM scores each pixel in the original image by its importance for a CNN's class prediction. The Grad-CAM provides coarse pixel-level class activation maps for each image which are scaled from 0 to 1 and multiplied by their HTT confidence scores for stability.

(3) Inter-HTT Adjustments. The original ADP database has no non-tissue labels, so *background* maps must be produced for both morphological and functional modes; ADP also omits non-functional labels for the functional mode, so *other* maps must also be produced. This allows HistoSegNet to avoid making predictions where no valid pixel class from ADP exists. The *background* activation is assumed to be regions of high white illumination which are not transparent-staining tissues (e.g. white/brown adipose, glandular/transport vessels); it is generated by applying a scaled-and-shifted sigmoid to the mean-RGB image, then subtracting the transparent-staining class activations, and applying a 2D Gaussian blur. The *other* activation is assumed to be regions of low activation for the *background* and all other functional tissues; it is generated by taking the 2D maximum of: (1) all other functional type activations, (2) white and brown adipose activations (from the morphological mode), and (3) the background activation. Then, this probability map is subtracted from one and scaled by 0.05. Finally, overlapping Grad-CAMs are differentiated by subtracting each activation map from the 2D maximum of the other Grad-CAMs - in locations of overlap, this suppresses weak activations overlapping with strong activations and improves results for dense CRF.

(4) Dense CRF. The resultant activation maps are still coarse and poorly conform to object contours, so the dense Conditional Random Field (CRF) (Krähenbühl and Koltun, 2011) is used, with an appearance kernel and a smoothness kernel being applied for 5 iterations using different settings each for the morphological and functional modes.

Ablative Study. SEC and DSRG use VGG16 to generate weak localization seeds while HistoSegNet uses a much shallower 3-block VGG16 variant; this raises the question, "Is network architecture important for WSSS performance?" This issue has never been explored before, so to answer this, we analyzed the performance of HistoSegNet using eight variant architectures of VGG16, named M1 (i.e. VGG16), M2, M3, M4, M5, M6, M7, and X1.7 (i.e. the one used in HistoSegNet) (see Figure 6). M1 through M4 analyze the effect of *network depth*: they all use GAP for vectorization and a single fully-connected layer, but differ in the number of convolutional blocks: 5, 4, 3, and 2 respectively. M5 through M7, on the other hand, analyze the effect of the *vectorization operation*: they all have 3 convolutional blocks and a single fully-connected layer, but use GAP, Flatten, and GMP for vectorization respectively. Finally, X1.7 analyzes the effect of *hierarchical binary relevance* (HBR) (Tsoumakas et al., 2009): it is identical to M7 but trains on all 51 classes of the ADP class set and tests on only the 31 segmentation classes. All eight networks were trained on Keras (TensorFlow backend) for 80 epochs with cyclical learning rate and a batch size of 16; they were evaluated for classification on the test set and for segmentation on the *segtest* set (both ADP-morph and ADP-func).

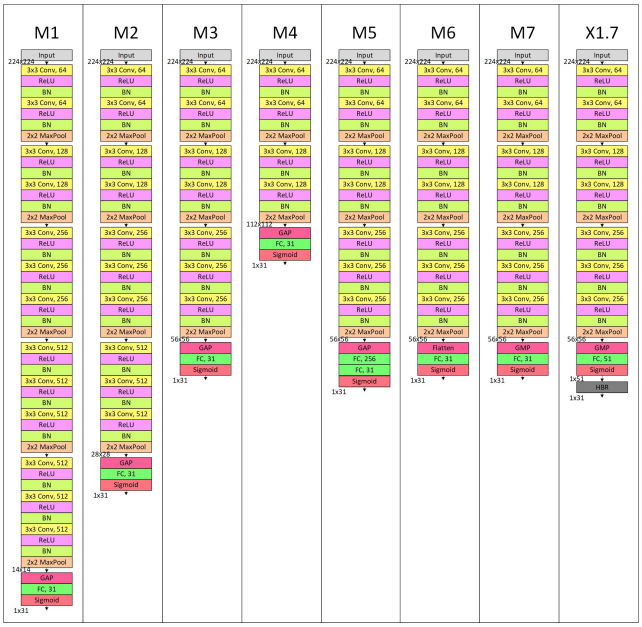
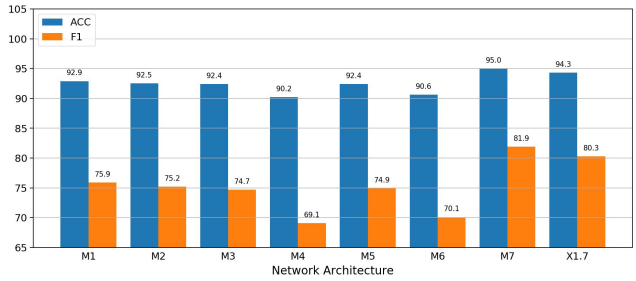
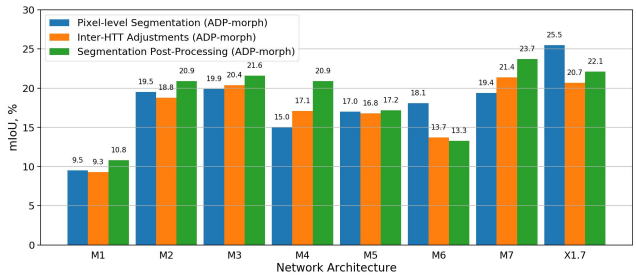


Fig. 6 Overview of the eight ablative architectures used to study the effect of network architecture on WSSS performance.

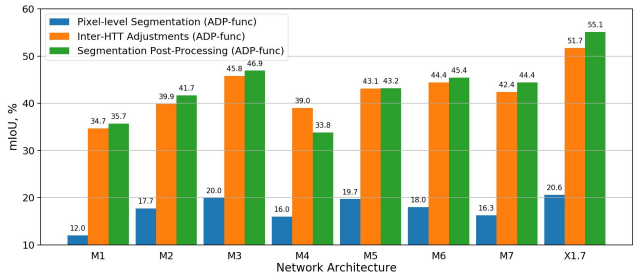
For classification (Figure 7(a)), networks with greater depth (i.e. M1) predict better than those with lesser depth (i.e. M2-M4), networks vectorized with GMP (i.e. M7) predict better than with GAP (i.e. M5) and Flatten (i.e. M6), and networks without HBR (i.e. M7) predict better than those with HBR (i.e. X1.7). But for segmentation, a different pattern emerges. For the morphological types (Figure 7(b)), although GMP vectorization and no HBR (i.e. M7) are still superior, lesser depth is beneficial up to 3 blocks (i.e. M3). For the functional types (Figure 7(c)), lesser depth is also beneficial up to 3 blocks (i.e. M3), but Flatten vectorization (i.e. M6) and with HBR (i.e. X1.7) are superior in this case. These results show that the classification network design is important for subsequent WSSS performance and that deeper networks such as VGG16 may perform well on classification but fail on segmentation due to their smaller convolutional feature maps.



(a) Classification performance



(b) Segmentation performance (ADP-morph)



(c) Segmentation performance (ADP-func)

Fig. 7 Performance of the eight ablative architectures in (a) classification, (b) morphological segmentation, and (c) functional segmentation. Deeper networks (i.e. M1) perform better than shallow networks (i.e. M2-M4) and vectorizing with GMP (i.e. M7) is better than GAP or Flatten (i.e. M5, M6) for classification, but shallower networks are better for segmentation up to 3 blocks (i.e. M3). Note that some scales do not start at zero.

5 Performance Evaluation

In this section, the four state-of-the-art methods are modified for the three representative segmentation datasets and their relative performance is evaluated. Until this point, there have been few attempts to apply WSSS methods to different image domains: SEC, DSRG, and IRNet have been developed for PASCAL VOC2012, while HistoSegNet has been developed for ADP. Hence, it is imperative to assess whether certain methods out-perform others on different segmentation datasets.

5.1 Setup

The original WSSS methods were developed to use different classification CNN architectures: VGG16 for SEC and DSRG, the shallower X1.7 for HistoSegNet, and the deeper ResNet-50 for IRNet. To avoid the possibility that the classification CNN choice would unfairly favour certain methods over others, we chose to implement all four WSSS methods with each of VGG16 and X1.7 - eight network-method configurations result. As Hierarchical Binary Relevance (HBR) was used in X1.7 to leverage the hierarchical class taxonomy in ADP, it is omitted for the non-hierarchical datasets (PASCAL VOC2012 and DeepGlobe) and denoted as M7. To generate seeds for the self-supervised methods, we decided to select confidence thresholds for SEC and DSRG that ensured less than 50% of the training set images were covered by seeds, which was heuristically determined to be optimal (this will be covered in more detail in Section 6.2). For IRNet, confidence thresholds were tuned using coordinate descent.

In common with the original practice used in SEC, DSRG, and HistoSegNet, images were first resized to 321×321 and 224×224 for VGG16 and M7 respectively, with flipping and moderate scaling used as image augmentation. We neither explored other image re-sizing techniques nor different receptive fields for the sake of simplicity. All CNNs were trained to convergence in 80 epochs with cyclical learning rate (Smith, 2017) (triangular policy, between 0.001 and 0.02 with a period of 10 epochs and 0.5 decay every 20 epochs). For SEC and DSRG, a simple stepwise decaying learning rate was used, starting at 0.0001 and 0.5 decay every 4 epochs. A constant learning rate of 0.1 and weight decay of 0.0001 was used for training IRNet for 3 epochs, following the authors' settings. All trainable weights in the CNNs, SEC, DSRG, and IRNet were pre-initialized initialized from ImageNet; this improved performance, even for ADP and DeepGlobe. See Section 2 of the Supplementary Materials for the CNN training details, Section 5 for SEC and DSRG.

Furthermore, non-foreground objects (e.g. background, other, are handled differently by all four methods, so the same approach is used for each dataset in all four methods to ensure fair comparison. We modified the openly-available Tensorflow implementations of SEC[1] and DSRG[2], as well as the Keras implementation of HistoSegNet[3]. We have released the full evaluation code for this paper online[4].

For each dataset, the eight network-method configurations are quantitatively ranked against the ground-truth annotated evaluation sets using the mean Intersection-over-Union (mIoU) metric, which measures the percent overlap between predicted (P) and ground-truth segmentation masks (T), av-

[1] https://github.com/xtudbxk/SEC-tensorflow
[2] https://github.com/xtudbxk/DSRG-tensorflow
[3] https://github.com/lyndonchan/hsn_v1
[4] https://github.com/lyndonchan/wsss-analysis

eraged across all C classes (see Equation 1). Qualitative evaluation is provided by visual inspection of the segmentation quality. Grad-CAM (using the most confident class at each pixel) is used as the baseline for both qualitative and quantitative evaluations (see Section 3 of the Supplementary Materials for details).

$$\text{mIoU} = \frac{1}{C}\sum_{c=1}^{C}\frac{|P_c \cap T_c|}{|P_c \cup T_c|} \tag{1}$$

5.2 Atlas of Digital Pathology (ADP)

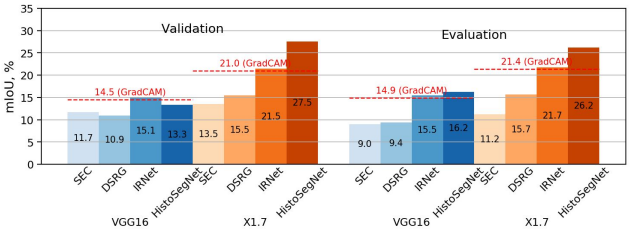
HistoSegNet was originally developed for ADP and hence needs no modifications, but SEC, DSRG, and IRNet were modified to generate *background* and *other* functional class activations by measuring the white level and the negative of the maximum of other functional classes. The foreground CAMs were thresholded at 90% of the maximum value for SEC and DSRG, with the same overlap strategy used; thresholding at 0.9 was used for IRNet. SEC and DSRG were trained for 8 epochs, IRNet for 3 epochs. See Sections 4.3, 4.4 of the Supplementary Materials for the detailed settings and training progress of SEC and DSRG in ADP-morph and ADP-func respectively; see Section 5.3, and 5.4 for IRNet.

Quantitative Performance

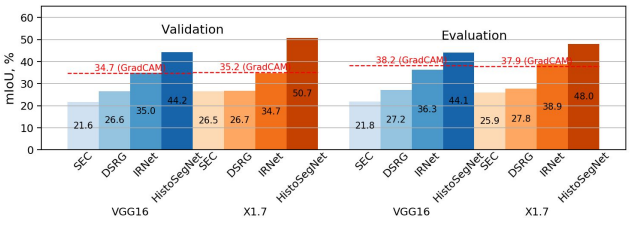
When assessed against the ground-truth evaluation set for both morphological and functional types (see Figure 8), it may be seen that (1) HistoSegNet is the only method that consistently out-performs the baseline Grad-CAM and that (2) the X1.7 network (which was designed for ADP) is superior to VGG16. Among the self-supervised methods, SEC performs worst, followed by DSRG; IRNet performs as well as Grad-CAM. As HistoSegNet was tuned with the validation set, it performs somewhat worse on the evaluation set (which has fewer unique classes per image).

Qualitative Performance

Figure 9 visualizes the segmentation performances for select patches. For the morphological types (see Figure 9(b)), the X1.7 configurations are superior to the VGG16 configurations (since the X1.7 Grad-CAMs correspond better to the smaller segments). While SEC and DSRG correspond well with object contours, they tend to over-exaggerate object sizes whereas HistoSegNet does not. For example, in image (1) of 9(b), only X1.7-HistoSegNet accurately segments the *simple cuboidal epithelium* of the thyroid glands (in green), although it struggles to delineate the *lymphocytes* (purple) in image (4) and *neuropil* (blue) in image (6). Similar behaviour is observed for the functional types (see Figure 9(c)): in images (1)-(4), only HistoSegNet detects small transport vessels (in fuchsia) although it produces false positives in images (5)-(6).

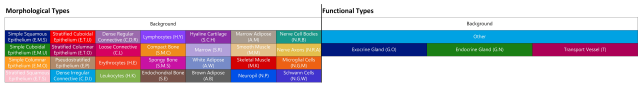


(a) Morphological types

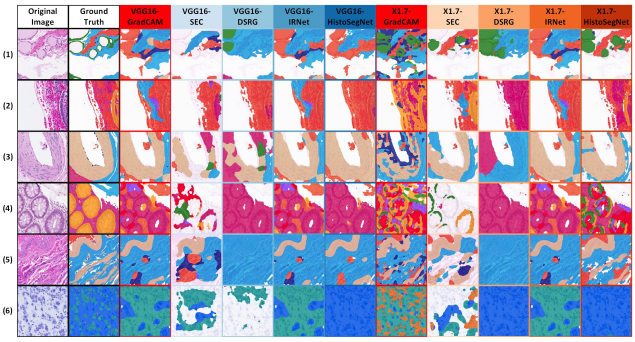


(b) Functional types

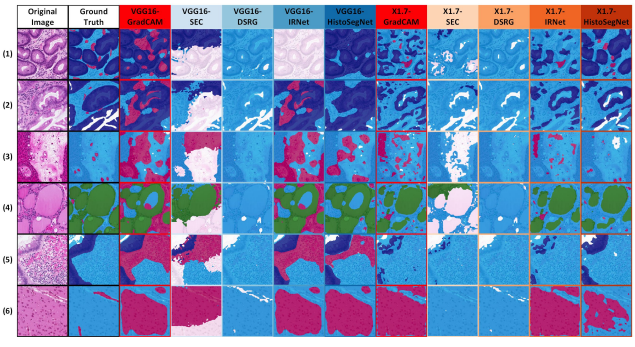
Fig. 8 ADP: quantitative performance of evaluated configurations (and baseline Grad-CAM) on the tuning (left) and evaluation set (right).



(a) Colour key



(b) Morphological types



(c) Functional types

Fig. 9 ADP: qualitative performance of evaluated configurations (and baseline Grad-CAM), on select evaluation patch images.

5.3 PASCAL VOC2012

SEC, DSRG, and IRNet were originally developed for PASCAL VOC2012 but new seeds were generated using our experimental framework and utilized for all methods. We used the *background* activation from SEC (i.e. the negative class sum of CAMs from the "background" network) for all four methods (including HistoSegNet) since the DRFI (Jiang et al., 2013) from DSRG had no readily implementable code and HistoSegNet's white-illumination assumption is not applicable here. Both SEC and DSRG were trained for 16 epochs, IRNet for 3 epochs. See Sections 4.5 and 5.5 of the Supplementary Materials for the detailed settings and training progress of SEC and DSRG, and IRNet respectively.

Quantitative Performance

When assessed against the ground-truth evaluation set (see Figure 10), it may be seen that (1) only SEC and DSRG consistently out-perform the baseline Grad-CAM, with SEC being clearly superior and that (2) the VGG16 network is overall superior to the M7 network. SEC using M7 cues performs the best overall (slightly better than SEC with VGG16 cues). Furthermore, we obtained results for SEC, DSRG, and IRNet somewhat inferior to those originally reported. We suspect that (1) neglecting to use DRFI for background cue generation in DSRG and (2) minor implementation differences between the Caffe and TensorFlow implementations are responsible for this in SEC and DSRG, since we observed discrepancies between our generated cues and those provided by the authors. For IRNet, using square image resizing and shallower networks than ResNet-50 may have caused decreased performance.

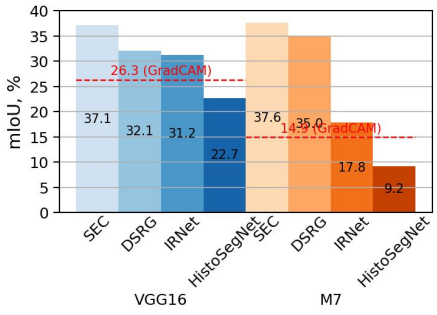
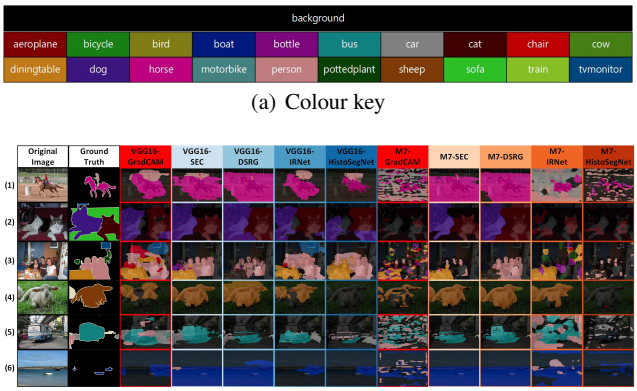


Fig. 10 PASCAL VOC2012: quantitative performance of evaluated network-method configurations (and baseline Grad-CAM) on the evaluation set.

Qualitative Performance

Figure 11 visualizes each configuration's segmentation results for several representative images. It is evident that the VGG16 Grad-CAM captures entire objects while the M7 Grad-CAM only captures parts and this results in the VGG16 configurations performing better. Furthermore, SEC and DSRG are able to correct mistakes in the original Grad-CAM (possibly due to the seeding loss function being well-suited to this dataset) whereas HistoSegNet often connects Grad-CAM segments to the wrong objects. In image (3), VGG16-HistoSegNet confuses the *diningtable* segment (yellow) with *person* segments (peach) while M7-HistoSegNet only segments heads and arms as *person*. All methods struggle most to differentiate objects that frequently occur together, such as *boat* and water in image (6).



(a) Colour key

(b) Segmentation results

Fig. 11 PASCAL VOC2012: qualitative performance of evaluated configurations (and baseline Grad-CAM), on select evaluation images

5.4 DeepGlobe Land Cover Classification

The DeepGlobe Land Cover Classification dataset was intended for fully-supervised semantic segmentation, so no published WSSS has ever been developed for it. We ignore the extremely uncommon *unknown* class for non-land cover objects, so all four methods consider the six land cover classes to be foreground. SEC and DSRG were trained for 13 epochs, IRNet for 3 epochs. See Sections 4.6 and 5.6 of the Supplementary Materials for the detailed settings and training progress of SEC and DSRG, and IRNet respectively.

Quantitative Performance

When assessed against the ground-truth evaluation set (see Figure 12), (1) only DSRG and IRNet consistently outperform the baseline Grad-CAM, although DSRG appears superior in general and (2) the M7 network is superior to VGG16, despite M7's Grad-CAM being inferior. DSRG using M7 cues performs the best overall. None of the methods were developed for DeepGlobe and the best fully-supervised method, DFCNet (Tian et al., 2019), which attained a 52.24% mIoU on the unreleased validation set, performs far better. Nonetheless, DSRG, IRNet, and HistoSegNet all out-perform the baseline while SEC does not.

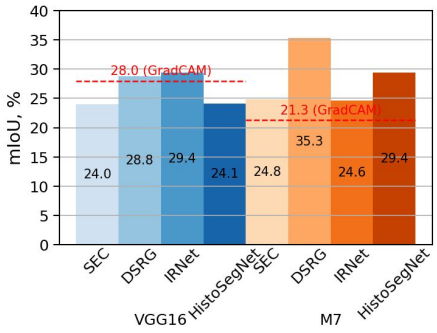
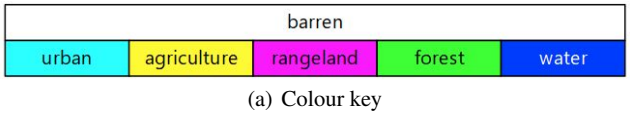


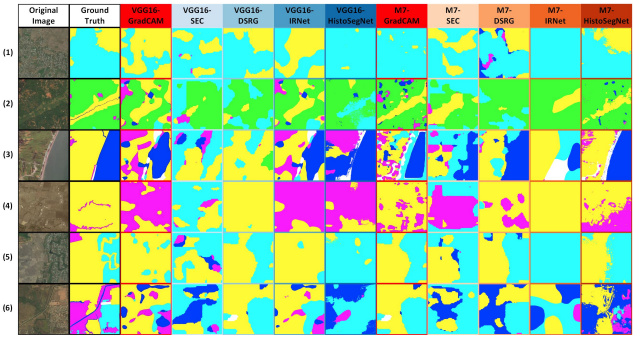
Fig. 12 DeepGlobe: quantitative performance of evaluated configurations (and baseline Grad-CAM) on the evaluation set.

Qualitative Performance

Figure 13 displays the segmentation results for several images. Visually, all four methods predict rather similarly, although DSRG and HistoSegNet capture small details better, and VGG16 methods tend to produce coarser predictions than M7. Unlike in VOC2012, the Grad-CAMs already capture the rough locations of the segments accurately, and only minor modifications are needed. For example, the M7 Grad-CAMs successfully detect the *agriculture* segment (yellow) in the middle of image (2) and the *rangeland* (magenta) in the bottom of image (4) but only HistoSegNet retains these preliminary segments. All methods struggle with segmenting *water* (blue), however, as shown in image (6).



(a) Colour key



(b) Segmentation results

Fig. 13 DeepGlobe: qualitative performance of evaluated configurations (and baseline Grad-CAM), on select evaluation images.

6 Analysis

Since the same four WSSS methods were compared with identical classification networks (or the closest equivalents) with the same evaluation setup in three datasets, it is possible to compare their comparative suitability for each dataset and observe some common themes. This is crucial, since WSSS in other image domains than natural scene images and histopathology images has been largely unexplored and applying them to these image domains requires an understanding of which approaches are best suited to the dataset at hand even before training. In this section, we analyze (1) the effect of the sparseness of classification network cues, (2) whether self-supervised learning is beneficial, and (3) how to address high class co-occurrence in the training set.

6.1 Effect of Classification Net Cue Sparseness

In most WSSS methods, not much attention is paid to the design of the classification network used. SEC and DSRG use VGG16 and HistoSegNet uses X1.7 (or M7). However, our experimental results showed that the choice of classification network has a significant effect on subsequent WSSS performance. Heuristically, we observed that networks generating sparser Grad-CAM segments would also perform better on datasets with more ground-truth segments. This was true for both the baseline Grad-CAMs and also subsequent WSSS performance. In Figure 14, this is demonstrated using a sample image from VOC2012 and ADP-func: the selected VOC2012 image has three ground-truth segments, while the ADP-func image has eight. VGG16's Grad-CAM predicts fewer segments because its final feature map is sized 41×41 (with input size of 321×321), but predicts sparser cues with M7 (and X1.7) because its final feature map is 56×56 (with input size of 224×224) and hence requires less upsampling. While VGG16 captures the spatial extent of the *person* and *horse* better than M7 in VOC2012, it is too coarse for ADP-func and X1.7 performs better.

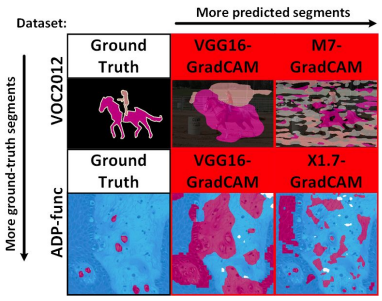


Fig. 14 Networks with more predicted segments (i.e. X1.7/M7) perform better than those with fewer (VGG16) on datasets with more segments (ADP-func) than fewer (VGG16).

This heuristic observation is also confirmed by quantitative analysis of the relation between the number of ground-truth instances and segmentation performance in the evaluation set. In Figure 15, the evaluation set mIoU of each configuration is shown for the three datasets after ordering by increasing number of ground-truth instances. VGG16 configurations (in shades of blue) perform best in datasets with fewer ground-truth instances ($\leq$1.65), while M7 configurations (in shades of orange) perform best in datasets with more ground-truth instances ($\geq$1.68). The effect is not insignificant, causing a mean difference of 5.22% in mIoU, and is especially pronounced for HistoSegNet. These results suggest that it is worthwhile to select a classification network with appropriately sparse cues for each new dataset based on the number of ground-truth instances.

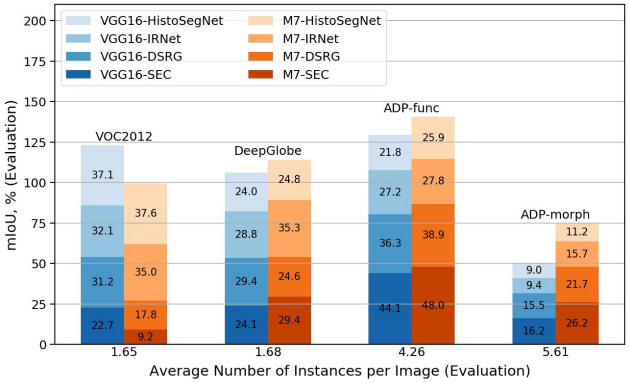


Fig. 15 Datasets with more ground-truth instances tend to be segmented better by X1.7/M7 (which has a larger feature map) than VGG16 (which has a smaller feature map).

6.2 Is Self-Supervised Learning Beneficial?

The prevailing approach to WSSS is currently to generate weak cues using CAM or Grad-CAM for self-supervised learning of an FCN (as used by SEC and DSRG). While this approach works well for natural scene images, our experimental results showed that it is clearly inferior for histopathology and of dubious value for satellite images. Why does self-supervised learning work well for some images and not others? Is it possible to determine ahead of time which approach is suitable for a given dataset before training? Heuristically, it was observed that self-supervised learning performance was heavily dependent on the degree to which ground-truth segments were already covered by the thresholded Grad-CAM seeds (adjusted to cover just under 50% of the image). In Figure 11, a sample image and the associated M7/X1.7 Grad-CAM segmentation is shown from VOC2012 and ADP-func respectively. The M7/X1.7 cue covers very little of the ground-truth *sheep* (tan-coloured) in the VOC2012 image but covers almost the entire ground-truth *other* in the ADP-func image; the self-supervised methods (SEC, DSRG, and IRNet) subsequently segment the VOC2012 image better while HistoSegNet (which is not self-supervised) segments the ADP-func image better.

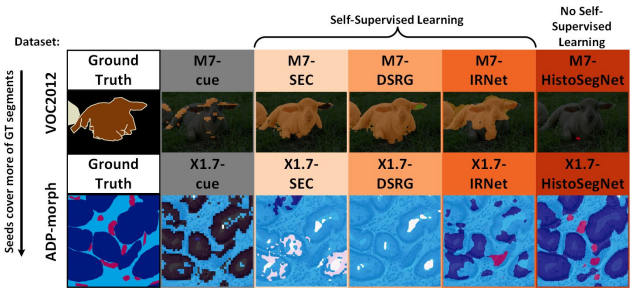


Fig. 16 Methods without self-supervised learning (i.e. HistoSegNet) perform better on datasets where seeds already cover much of the ground-truth segments (i.e. ADP-func), known as mean recall. Self-supervised learning seems to be a poor choice in these cases.

Quantitatively, the degree of overlap between predicted seeds and ground-truth segments in the evaluation set can be measured by mean recall, which is the percent of ground-truth pixels T_c that are predicted P_c correctly for each class c, then averaged across all C classes (see Equation 2):

$$\text{Recall} = \frac{1}{C}\sum_{c=1}^{C}\frac{|P_c \cap T_c|}{|T_c|} \tag{2}$$

In Figure 17, the evaluation set mIoU of each configuration is shown for each dataset's cues after ordering by increasing mean recall. Self-supervised methods (SEC, DSRG, and IRNet) tend to perform better than the cues for datasets with low seed coverage (such as VOC2012 and DeepGlobe) but HistoSegNet performs better for datasets with high seed coverage (ADP-func and ADP-morph). This suggests that, when applying WSSS to a new dataset, one should choose a self-supervised method (e.g. SEC and DSRG) if seed recall is low ($< 40\%$) and a method without self-supervised learning (e.g. HistoSegNet) if seed recall is high ($\geq 40\%$). This makes much intuitive sense, since CAM/Grad-CAM is notorious for only segmenting parts of ground-truth objects in VOC2012. Hence, self-supervised methods were developed with loss functions to encourage predicting liberally from minimal seeds by rewarding true positives and not penalizing false positives. While this strategy works well when seeds cover little of the ground-truth, it is clearly detrimental when the seeds cover much of the ground-truth.

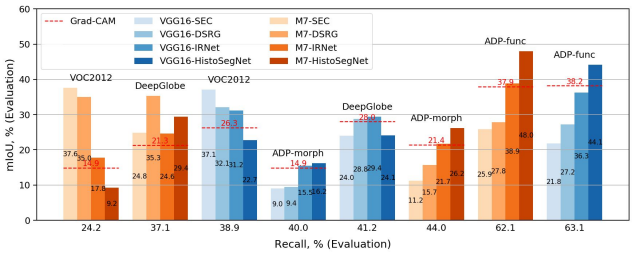
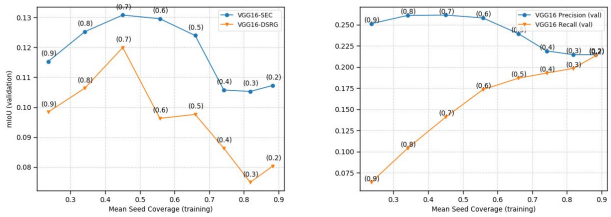


Fig. 17 Datasets with higher mean seed recall (after thresholding below 50% seed coverage) are overwhelmingly better segmented by non-self supervised learning methods (i.e. HistoSegNet). Where recall is low, self-supervised methods (i.e. SEC, DSRG, IRNet) are superior.

Although the previous analysis determined that high mean recall in the week seeds is beneficial for SEC and DSRG, lowering the seed threshold to ensure more than 50% of the image area is seeded (mean seed coverage) and thus increase the mean recall is not a feasible strategy either. In Figure 18(a), a simple ablative study is shown of the relationship between mIoU in ADP-morph for VGG16-SEC and VGG16-DSRG with different seed threshold levels (i.e. 20% to 90%). Figure 18(b) shows the effect of these same threshold levels on seed precision and recall. Although decreasing the seed threshold level to 20% (and increasing seed coverage to 88.3%) increases mean recall to 21.3%, the optimal seed threshold level for SEC and DSRG mIoU is actually when seed coverage is just below 50%, resulting in a lower mean recall of 19.3%. This simple analysis indicates that self-supervised methods such as SEC and DSRG are inherently ill-suited for datasets with low seed recall; the seed threshold level should be fixed so seed coverage is just below 50% and decreasing the threshold to increase recall cannot improve performance. Full details are available in Section 9 of the Supplementary Materials.

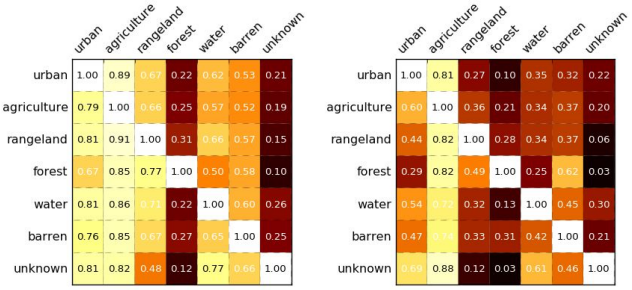


(a) SEC/DSRG mIoU. (b) VGG16 seed precision/recall

Fig. 18 Mean seed coverage in the training set (threshold levels in brackets), plotted versus mIoU of VGG16-SEC and VGG16-DSRG for ADP-morph validation set (left), versus seed precision/recall (right). Although mean recall is maximized by increasing seed coverage, SEC and DSRG perform best when seed coverage is fixed at just below 50%.

6.3 Addressing High Class Co-Occurrence

Learning semantic segmentation from image labels is the weakest form of supervision possible because it provides no location information of the objects. This information must be inferred by their presence or absence in the annotated images. Logically, it would make sense that image label supervision would be least informative in datasets where the classes frequently occur together. In the extreme case that two labels always occur together, it would be impossible to learn to spatially separate them. The DeepGlobe dataset, for example, has very high levels of class co-occurrence (see Figure 19) - the classes in the original training set (see Figure 19(a)) regularly co-occur in more than 50% of images (except for *forest* and *unknown*). To assess whether simply reducing class co-occurrence would improve WSSS performance, we removed half of these original training images with the most class labels (defined as the sum of overall class counts for each image) and then retrained - we call this process "balancing" the class distribution. As a result, the class co-occurrence is significantly reduced (see Figure 19(b)) in all classes except *urban* and *agriculture*.



(a) Training set, without balancing (75% train) (b) Training set, with balancing (37.5% train)

Fig. 19 Normalized class co-occurrences in the ground-truth image annotations of different DeepGlobe train-test splits. By removing the training images with the most annotated classes (i.e. balancing), training set class co-occurrence is significantly reduced in the *rangeland*, *forest*, and *water* classes.

When we use these two different train-test splits and evaluate on the same test set, we obtain the quantitative mIoU performances shown in Figure 20 - results before balancing are shown on the left and after balancing on the right. Although performance in certain methods deteriorate significantly after balancing, the best performance improves in the classes that experienced the greatest changes from the balancing, such as *agriculture*, *forest*, and *water*. See Sections 4.7 and 5.7 of the Supplementary Materials for the detailed training progress of SEC and DSRG, and IRNet respectively.

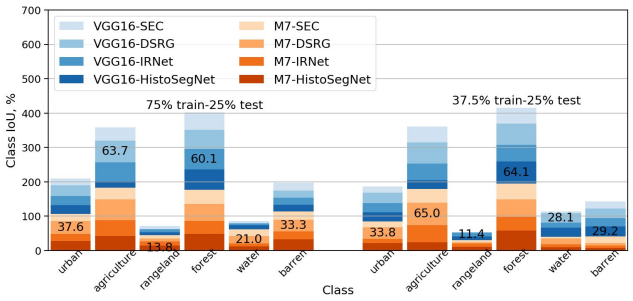


Fig. 20 Class IoU of the best-performing configuration increases for the *agriculture*, *forest*, and *water* after balancing (right), compared to before balancing (left). Some classes experience a performance decrease, suggesting that the balancing method has room for improvement.

This is confirmed upon inspecting some segmentation results for images containing *forest* and *water*, as shown in Figure 21 for VGG16-HistoSegNet. Since balancing drastically reduces class co-occurrence in these two classes, VGG16-HistoSegNet learns to delineate *forest* from *agriculture* better in image (a) and associate *water* with the river on the left of image (b). This shows that class co-occurrence is a significant challenge for WSSS from image labels but a simple technique to reduce it can help performance in the affected classes. Overall, the mean mIoU decreases from 27.5% to 26.5% after balancing, so we hypothesize that more effective methods of reducing class co-occurrence in the training set can more robustly improve WSSS performance.

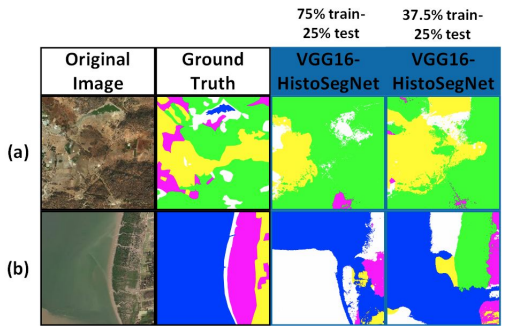


Fig. 21 Segmentation by VGG16-HistoSegNet for DeepGlobe, trained without (second from right) and with class balancing (right). Performance improves most in classes experiencing the greatest changes, such as *forest* and *water*.

7 Conclusion

Weak supervision with image labels is a promising approach to semantic segmentation (WSSS) because image annotations require significantly less expense and time than the pixel annotations needed for full supervision. To date, state-of-the-art WSSS methods have built their methodologies exclusively around natural scene images. However, the lack of methods built for alternative image domains indicates there is an implicit assumption that these methodologies are generalizable with minor modifications. However, while the major remaining challenges for natural scene images concern separating background from foreground and segmenting entire objects instead of parts, alternative image domains such as histopathology and satellite images present different challenges, such as ambiguous boundaries and class co-occurrence. This paper is the first to analyze whether state-of-the-art methods developed for natural scene images still perform acceptably on histopathology and satellite images and compares their performances against a method developed for histopathology images.

Our experiments indicated that state-of-the-art methods developed for natural scene (i.e. SEC and DSRG) and histopathology images (i.e. HistoSegNet) indeed performed best in their intended domains. Furthermore, we showed that most methods perform moderately well for satellite images. Many methods performed poorly on datasets they were not designed to solve (such as HistoSegNet in VOC2012, SEC and DSRG in ADP), although IRNet seemed to be most robust overal to dataset choice. We found that the sparseness of a classification network's baseline Grad-CAM had a significant effect on subsequent segmentation performance if the ground-truth segments were also sparse. We also observed that the self-supervised learning approach to WSSS was only beneficial if the optimally-seeded cues covered little of the ground-truth segments (low mean recall), and that methods forgoing the self-supervised learning approach performed better otherwise. Finally, we demonstrated the negative effect of class co-occurrence on segmentation performance and showed that even a simple method of reducing class co-occurrence can alleviate this problem.

The findings of our paper clearly indicate that mainstream methodologies are poorly suited for these other image domains. We believe that more work is needed to develop alternative methodologies for WSSS which are either specialized for these image domains or are at least more generalizable to them. Instead of the current focus on improving the recall of activation map seeds, perhaps devising new loss functions which refine ambiguous activation map boundaries or address high class co-occurrence would be better directions to explore in the future for histopathology and satellite images. Given the ease of collecting weak annotations and the inability of state-of-the-art algorithms to properly segment images from alternative image domains, the authors believe that it is imperative that more work be done to develop new methods capable of generalizing to different image domains.

References

Ahn J, Kwak S (2018) Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. *CoRR* abs/1803.10464, 1803.10464

Ahn J, Cho S, Kwak S (2019) Weakly supervised learning of instance segmentation with inter-pixel relations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2209–2218

Aresta G, etal (2018) BACH: grand challenge on breast cancer histology images. *CoRR* abs/1808.04277, 1808.04277

Audebert N, Le Saux B, Lefèvre S (2017) Segment-before-detect: Vehicle detection and classification through semantic segmentation of aerial images. *Remote Sensing* 9(4):368

Bearman A, Russakovsky O, Ferrari V, Fei-Fei L (2016) Whats the point: Semantic segmentation with point supervision. In: European Conference on Computer Vision, Springer, pp 549–565

Beck AH, Sangoi AR, Leung S, Marinelli RJ, Nielsen TO, Van De Vijver MJ, West RB, Van De Rijn M, Koller D (2011) Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Science translational medicine* 3(108):108ra113–108ra113

Brostow GJ, Shotton J, Fauqueur J, Cipolla R (2008) Segmentation and recognition using structure from motion point clouds. In: European Conference on Computer Vision, Springer, pp 44–57

Caesar H, Uijlings J, Ferrari V (2018) Coco-stuff: Thing and stuff classes in context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1209–1218

Chan L, Hosseini MS, Rowsell C, Plataniotis KN, Damaskinos S (2019) Histosegnet: Semantic segmentation of histological tissue type in whole slide images. In: International Conference on Computer Vision (ICCV)

Chen H, Qi X, Yu L, Heng PA (2016) Dcan: Deep contour-aware networks for accurate gland segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2014) Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:14127062*

Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2017a) Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40(4):834–848

Chen LC, Papandreou G, Schroff F, Adam H (2017b) Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:170605587*

Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder-decoder with atrous separableconvolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 801–818

Ciresan D, Giusti A, Gambardella LM, Schmidhuber J (2012) Deep neural networks segment neuronal membranes in electron microscopy images. In: Advances in Neural Information Processing Systems 25, Curran Associates, Inc., pp 2843–2851

Cireşan DC, Giusti A, Gambardella LM, Schmidhuber J (2013) Mitosis detection in breast cancer histology images with deep neural networks. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013, Springer Berlin Heidelberg, Berlin, Heidelberg, pp 411–418

Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S, Schiele B (2016) The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3213–3223

Dai J, He K, Sun J (2015) Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp 1635–1643

Demir I, Koperski K, Lindenbaum D, Pang G, Huang J, Basu S, Hughes F, Tuia D, Raskar R (2018) Deepglobe 2018: A challenge to parse the earth through satellite images. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops

Durand T, Mordan T, Thome N, Cord M (2017) Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 642–651

Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A (2010) The pascal visual object classes (voc) challenge. *International Journal of Computer Vision* 88(2):303–338

Farabet C, Couprie C, Najman L, LeCun Y (2012) Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(8):1915–1929

Fu J, Liu J, Tian H, Li Y, Bao Y, Fang Z, Lu H (2019) Dual attention network for scene segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3146–3154

Gao J, Liao W, Nuyttens D, Lootens P, Vangeyte J, Pižurica A, He Y, Pieters JG (2018) Fusion of pixel and object-based features for weed mapping using unmanned aerial vehicle imagery. *International journal of applied earth observation and geoinformation* 67:43–53

Hariharan B, Arbelaez P, Bourdev L, Maji S, Malik J (2011) Semantic contours from inverse detectors. In: International Conference on Computer Vision (ICCV)

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778

Helber P, Bischke B, Dengel A, Borth D (2019) Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*

Hosseini MS, Chan L, Tse G, Tang M, Deng J, Norouzi S, Rowsell C, Plataniotis KN, Damaskinos S (2019) Atlas of digital pathology: A generalized hierarchical histological tissue type-annotated database for deep learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 11747–11756

Hou L, Samaras D, Kurc TM, Gao Y, Davis JE, Saltz JH (2016) Patch-based convolutional neural network for whole slide tissue image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2424–2433

Huang Z, Wang X, Wang J, Liu W, Wang J (2018) Weakly-supervised semantic segmentation network with deep seeded region growing. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp 7014–7023

Jia Z, Huang X, Eric I, Chang C, Xu Y (2017) Constrained deep weak supervision for histopathology image segmentation. *IEEE transactions on medical imaging* 36(11):2376–2388

Jiang H, Wang J, Yuan Z, Wu Y, Zheng N, Li S (2013) Salient object detection: A discriminative regional feature integration approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2083–2090

Kainz P, Pfeiffer M, Urschler M (2015) Semantic segmentation of colon glands with deep convolutional neural networks and total variation segmentation. *CoRR* abs/1511.06919, 1511.06919

Kather JN, Weis CA, Bianconi F, Melchers SM, Schad LR, Gaiser T, Marx A, Zöllner FG (2016) Multi-class texture analysis in colorectal cancer histology. *Scientific reports* 6:27988

Kather JN, Krisam J, Charoentong P, Luedde T, Herpel E, Weis CA, Gaiser T, Marx A, Valous NA, Ferber D, et al. (2019) Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS medicine* 16(1):e1002730

Kervadec H, Dolz J, Tang M, Granger E, Boykov Y, Ayed IB (2019) Constrained-cnn losses for weakly supervised segmentation. *Medical Image Analysis* 54:88–99

Kolesnikov A, Lampert CH (2016a) Improving weakly-supervised object localization by micro-annotation. *arXiv preprint arXiv:160505538*

Kolesnikov A, Lampert CH (2016b) Seed, expand and constrain: Three principles for weakly-supervised image segmentation. *CoRR* abs/1603.06098, 1603.06098

Kothari S, Phan JH, Young AN, Wang MD (2013) Histological image classification using biologically interpretable shape-based features. *BMC medical imaging* 13(1):9

Krähenbühl P, Koltun V (2011) Efficient inference in fully connected crfs with gaussian edge potentials. In: Advances in Neural Information Processing Systems, pp 109–117

Kumar N, Verma R, Sharma S, Bhargava S, Vahadane A, Sethi A (2017) A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE Transactions on Medical Imaging* 36(7):1550–1560, DOI 10.1109/TMI.2017.2677499

Kuo TS, Tseng KS, Yan JW, Liu YC, Wang YCF (2018) Deep aggregation net for land cover classification. In: CVPR Workshops, pp 252–256

Kwak S, Hong S, Han B (2017) Weakly supervised semantic segmentation using superpixel pooling network. In: Thirty-First AAAI Conference on Artificial Intelligence

Lee J, Kim E, Lee S, Lee J, Yoon S (2019) Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 5267–5276

Lenz M, Roumans NJ, Vink RG, van Baak MA, Mariman EC, Arts IC, de Kok TM, Ertaylan G (2016) Estimating real cell size distribution from cross-section microscopy imaging. *Bioinformatics* 32(17):i396–i404

Li W, Manivannan S, Akbar S, Zhang J, Trucco E, McKenna SJ (2016) Gland segmentation in colon histology images using hand-crafted features and convolutional neural networks. In: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), pp 1405–1408, DOI 10.1109/ISBI.2016.7493530

Lin D, Dai J, Jia J, He K, Sun J (2016) Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3159–3167

Lin H, Chen H, Dou Q, Wang L, Qin J, Heng P (2017a) Scannet: A fast and dense scanning framework for metastatic breast cancer detection from whole-slide images. *CoRR* abs/1707.09597, 1707.09597

Lin H, Chen H, Dou Q, Wang L, Qin J, Heng PA (2018) Scannet: A fast and dense scanning framework for metastastic breast cancer detection from whole-slide image. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, pp 539–546

Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: Common objects in context. In: European Conference on Computer Vision, Springer, pp 740–755

Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017b) Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2117–2125

Liu C, Yuen J, Torralba A (2010) Sift flow: Dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(5):978–994

Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3431–3440

Malon C, Cosatto E (2013) Classification of mitotic figures with convolutional neural networks and seeded blob features. In: Journal of Pathology Informatics

Mnih V, Hinton GE (2010) Learning to detect roads in high-resolution aerial images. In: European Conference on Computer Vision, Springer, pp 210–223

Mottaghi R, Chen X, Liu X, Cho NG, Lee SW, Fidler S, Urtasun R, Yuille A (2014) The role of context for object detection and semantic segmentation in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Neuhold G, Ollmann T, Rota Bulo S, Kontschieder P (2017) The mapillary vistas dataset for semantic understanding of street scenes. In: Proceedings of the IEEE International Conference on Computer Vision, pp 4990–4999

Nivaggioli A, Randrianarivo H (2019) Weakly supervised semantic segmentation of satellite images. *arXiv preprint arXiv:190403983*

Papandreou G, Chen L, Murphy K, Yuille AL (2015) Weakly- and semi-supervised learning of a DCNN for semantic image segmentation. *CoRR* abs/1502.02734, 1502.02734

Pathak D, Shelhamer E, Long J, Darrell T (2014) Fully convolutional multi-class multiple instance learning. *CoRR* abs/1412.7144, 1412.7144

Pathak D, Krahenbuhl P, Darrell T (2015) Constrained convolutional neural networks for weakly supervised segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp 1796–1804

Prince SJ (2012a) *Computer vision: models, learning, and inference*, Cambridge University Press, pp 201–208

Prince SJ (2012b) *Computer vision: models, learning, and inference*, Cambridge University Press, p 15

Rahnemoonfar M, Murphy R, Miquel MV, Dobbs D, Adams A (2018) Flooded area detection from uav images based on densely connected recurrent neural networks. In: IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium, IEEE, pp 1788–1791

Riordan DP, Varma S, West RB, Brown PO (2015) Automated analysis and classification of histological tissue features by multi-dimensional microscopic molecular profiling. *PloS one* 10(7):e0128975

Robinson C, Hou L, Malkin K, Soobitsky R, Czawlytko J, Dilkina B, Jojic N (2019) Large scale high-resolution land cover mapping with multi-resolution data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 12726–12735

Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, Springer, pp 234–241

Roux L, Racoceanu D, Loménie N, Kulikova M, Irshad H, Klossa J, Capron F, Genestie C, le Naour G, Gurcan MN (2013) Mitosis detection in breast cancer histological images an icpr 2012 contest. In: Journal of Pathology Informatics

Saleh F, Akbarian MSA, Salzmann M, Petersson L, Gould S, Alvarez JM (2016) Built-in foreground/background prior for weakly-supervised semantic segmentation. *CoRR* abs/1609.00446, 1609.00446

Seferbekov SS, Iglovikov V, Buslaev A, Shvets A (2018) Feature pyramid network for multi-class land segmentation. In: CVPR Workshops, pp 272–275

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp 618–626

Shapiro L, Stockman G (2000) *Computer Vision*, Pearson, pp 305–306

Shimoda W, Yanai K (2016) Distinct class-specific saliency maps for weakly supervised semantic segmentation. In: European Conference on Computer Vision, Springer, pp 218–234

Shkolyar A, Gefen A, Benayahu D, Greenspan H (2015) Automatic detection of cell divisions (mitosis) in live-imaging microscopy images using convolutional neural networks. *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* pp 743–746

Shotton J, Winn J, Rother C, Criminisi A (2006) Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In: European Conference on Computer Vision, Springer, pp 1–15

Sirinukunwattana K, Pluim JP, Chen H, Qi X, Heng PA, Guo YB, Wang LY, Matuszewski BJ, Bruni E, Sanchez U, et al. (2017) Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis* 35:489–502

Smith LN (2017) Cyclical learning rates for training neural networks. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, pp 464–472

Tian C, Li C, Shi J (2018) Dense fusion classmate network for land cover classification. In: CVPR Workshops, pp 192–196

Tian C, Li C, Shi J (2019) Dense fusion classmate network for land cover classification. *arXiv preprint arXiv:191108169*

Tsoumakas G, Katakis I, Vlahavas I (2009) Mining multi-label data. In: Data mining and knowledge discovery handbook, Springer, pp 667–685

Turkki R, Linder N, Kovanen P, Pellinen T, Lundin J (2016) Antibody-supervised deep learning for quantification of tumor-infiltrating immune cells in hematoxylin and eosin stained breast cancer samples. *Journal of Pathology Informatics* 7(1):38, DOI 10.4103/2153-3539.189703

Veta M, etal (2014) Assessment of algorithms for mitosis detection in breast cancer histopathology images. *Medical Image Analysis* DOI 10.1016/j.media.2014.11.010

Wang J, Sun K, Cheng T, Jiang B, Deng C, Zhao Y, Liu D, Mu Y, Tan M, Wang X, et al. (2019a) Deep high-resolution representation learning for visual recognition. *arXiv preprint arXiv:190807919*

Wang P, Huang X, Cheng X, Zhou D, Geng Q, Yang R (2019b) The apolloscape open dataset for autonomous driving and its application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*

Wang X, Chen H, Gan CHK, Lin H, Dou Q, Huang Q, Cai M, Heng PA (2018) Weakly supervised learning for whole slide lung cancer image classification. In: IEEE Transactions on Cybernetics

Wei Y, Feng J, Liang X, Cheng M, Zhao Y, Yan S (2017) Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. *CoRR* abs/1703.08448, 1703.08448

Wei Y, Xiao H, Shi H, Jie Z, Feng J, Huang TS (2018) Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 7268–7277

Xia F, Wang P, Chen X, Yuille AL (2017) Joint multi-person pose estimation and semantic part segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 6769–6778

Xie J, Liu R, Luttrell I, Zhang C, et al. (2019) Deep learning based analysis of histopathological images of breast cancer. *Frontiers in genetics* 10:80

Xu J, Schwing AG, Urtasun R (2015) Learning to segment under various forms of weak supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3781–3790

Xu J, Luo X, Wang G, Gilmore H, Madabhushi A (2016) A deep convolutional neural network for segmenting and classifying epithelial and stromal regions in histopathological images. *Neurocomputing* 191:214–223

Xu Y, Zhu JY, Eric I, Chang C, Lai M, Tu Z (2014) Weakly supervised histopathology cancer image segmentation and classification. *Medical image analysis* 18(3):591–604

Xu Y, Jia Z, Wang LB, Ai Y, Zhang F, Lai M, Eric I, Chang C (2017) Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. *BMC bioinformatics* 18(1):281

Yang Y, Newsam S (2010) Bag-of-visual-words and spatial extensions for land-use classification. In: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, pp 270–279

Yao X, Han J, Cheng G, Qian X, Guo L (2016) Semantic annotation of high-resolution satellite images via weakly supervised learning. *IEEE Transactions on Geoscience and Remote Sensing* 54(6):3660–3671

Ye L, Liu Z, Wang Y (2018) Learning semantic segmentation with diverse supervision. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, pp 1461–1469

Yu F, Xian W, Chen Y, Liu F, Liao M, Madhavan V, Darrell T (2018) Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:180504687*

Yuan Y, Chen X, Wang J (2019) Object-contextual representations for semantic segmentation. *arXiv preprint arXiv:190911065*

Zhang C, Li H, Wang X, Yang X (2015a) Cross-scene crowd counting via deep convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 833–841

Zhang C, Wei S, Ji S, Lu M (2019) Detecting large-scale urban land cover changes from very high resolution remote sensing images using cnn-based classification. *ISPRS International Journal of Geo-Information* 8(4):189

Zhang X, Su H, Yang L, Zhang S (2015b) Fine-grained histopathological image analysis via robust segmentation and large-scale retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 5361–5368

Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016) Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2921–2929

Zhou B, Zhao H, Puig X, Fidler S, Barriuso A, Torralba A (2017) Scene parsing through ade20k dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 633–641

Zhou Y, Zhu Y, Ye Q, Qiu Q, Jiao J (2018) Weakly supervised instance segmentation using class peak response. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3791–3800