A Brief Survey on Weakly Supervised Semantic Segmentation

https://doi.org/10.3991/ijoe.v18i10.31531

Youssef Ouassit[1](✉), Soufiane Ardchir[2], Mohammed Yassine El Ghoumari[2], Mohamed Azouazi[1]
[1] Faculty of sciences Ben M'sik, Casablanca, Morocco
[2] International school of marketing and management, Casablanca, Morocco
ouassit.youssef@gmail.com

Abstract—Semantic Segmentation is the process of assigning a label to every pixel in the image that share same semantic properties and stays a challenging task in computer vision. In recent years, and due to the large availability of training data the performance of semantic segmentation has been greatly improved by using deep learning techniques. A large number of novel methods have been proposed. However, in some crucial fields we can't assure sufficient data to learn a deep model and achieves high accuracy. This paper aims to provide a brief survey of research efforts on deep-learning-based semantic segmentation methods on limited labeled data and focus our survey on weakly-supervised methods. This survey is expected to familiarize readers with the progress and challenges of weakly supervised semantic segmentation research in the deep learning era and present several valuable growing research points in this field.

Keywords—deep learning, weakly supervision, semantic segmentation, computer vision

1 Introduction

Semantic segmentation is the task of assigning a semantic label to each pixel in an image, which is a fundamental task but still a challenging in computer vision field. As semantic segmentation can provide the information at the pixel level and then reduce the semantic gap between low level and high-level features, many real-world applications benefit from this task, such as self-driving vehicles, objects recognition, pedestrian detection, therapy planning, and computer-aided diagnosis. Semantic segmentation is distinguished from other computer vision tasks. In this context, in object classification the whole image is annotated with one or more semantic labels, in the object detection the system needs to know localization of the target objects in the scene, in semantic segmentation we answer both questions of what the object is and where is in the scene. Before the deep learning, many segmentation methods and algorithms have been proposed, such as methods based on the partial differential equations or statistics strategies. With sufficient training data, the supervised learning is able to considerably improve the capacity of a segmentation model. Recently, the deep learning techniques

has promoted semantic segmentation research. Fully Convolutional Network (FCN) [1], has dramatically increased the segmentation accuracy and paved the way for deep-learning-based semantic segmentation. Actually, many novel deep learning-based methods have been proposed and produce a high performance and remarkable improvement in effectiveness compared to the traditional methods. Actually, deep learning is the state-of-the-art in almost all public datasets.

Many survey papers on semantic segmentation has been proposed [2]–[9]. Most of them focus mainly on the traditional learning-based and fully supervised semantic segmentation methods, such as region proposal-based and CNN-based approaches, at the date of writing this survey no paper focus on weakly supervised methods in details. Our paper is different in the following aspects. First, in our paper we classify the methods based on different aspects, that is, the priors and hints used during the training process. Second, our paper particularly summarizes the methods focusing on weakly supervised segmentation methods, which is less reviewed and discussed in other surveys even if this field is greatly an active field of research. The organization of this survey is summarized as follows: Section 2 overviews in brief the task of semantic segmentation, and the common deep network architectures. Section 3 reviews the deep-learning-based semantic segmentation in weak supervision. In section 4 we present the common evaluation metrics used in semantic segmentation task. In Section 5 the most used loss functions are discussed. In Section 6, we introduce the commonly used datasets. In Section 7 we summarize the common challenges faced by the current methods and enumerate several growing research points and concludes our paper. Our expectation is that this survey helps researchers to become familiar with weakly supervised deep-learning-based for semantic segmentation from and provide some possible ideas and perspectives for a future works.

2 Overview

2.1 Semantic segmentation

As it's shown in Figure 1, Semantic segmentation is the process in which different parts of an image that belongs to the same object class are clustered together by assigning each pixel of an image a pre-defined section. The aim of semantic segmentation is to divide the image into exclusive subsets that simply represents the meaningful region of the original image. By comparing to other tasks in computer vision: The aim of image classification is to give labels of one or more categories to the whole image. The algorithm of image classification adds annotation to the image, such as a person, a cat, a dog, etc. But the object detection goes one step further by assigning a caption and localizing the object that detected in an image. For example, in Figure 1 Computer vision tasks, the objects located by object detection algorithm are highlighted with annotated rectangles. On the other hand, the aim of semantic segmentation is to highlights the object region and separate it from the background region. The algorithm accurately delineates the object boundaries in pixel level. This makes semantic classification a more challenging task than other computer vision task because it helps to completely reduce the semantic gap between high-level features and low-

level semantics. Recently, related to semantic segmentation new task have emerged as new research direction, panoptic segmentation, and instance segmentation. In instance segmentation each object is detected as an individual in the image with different categories labels, in panoptic segmentation we assign an instance label and a semantic label to each pixel. Difference between those tasks can be resumed in this way: traditional semantic segmentation focuses on "stuff", instance segmentation focus on "thing" and panoptic segmentation on "stuff+thing".

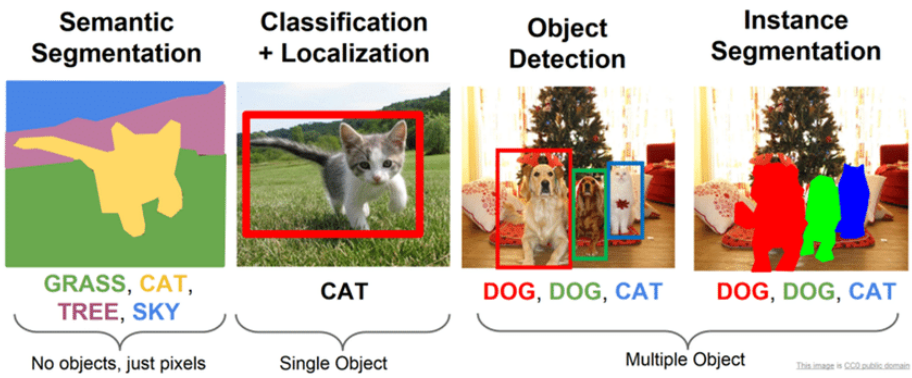

Fig. 1. Computer vision tasks

2.2 Deep learning for semantic segmentation

Recently, promising results have been achieved by deep learning methods in semantic segmentation task, generally, when training data and pixelwise labeling of images are sufficiently available, DNN are able to learn the mapping between the semantic label and its visual form. The process of learning reduces the gap between low-level features and high-level semantics, and makes the network more sensitive and aware to different semantic concepts. Next, the most common deep architecture that used in semantic segmentation will be reviewed in detail, as shown in Table 1.

FCN for semantic segmentation. Fully Convolution Neural Network was the first segmentation models that based on Convolution Neural Network and have high performance and remarkable accuracy in semantic segmentation task [1]. The proposed network contains many convolutional layers with one last up-sampling (deconvolving) layer at the end where the output is an activation maps and with the help of these maps the pixel-wise output can be calculated. In order to preserve the contextual spatial information within an image as the filtered input data penetrates deeper into the network, authors suggest fusing the output with shallower layers output. The fusion step is presented in Figure 2.

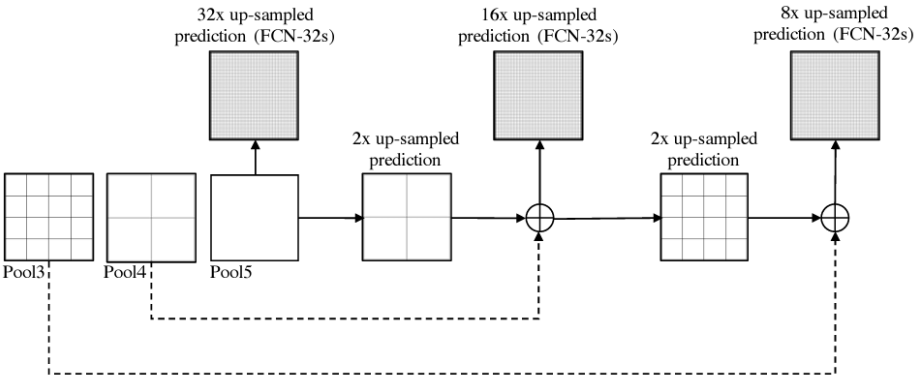

Fig. 2. Upsampling and fusion step of the fully convolution networks [1]

In FCN the main idea is to make the classical Convolution Neural Network to take arbitrary-sized images as input. This restriction of Convolution Neural Networks comes from the layers that are fully connected and fixed. In FCNs they only have pooling and convolutional layers that enable them to make accurate predictions on arbitrary sized inputs.

SegNet - a deep convolutional architecture (encoder and decoder). [10] It proposed a deep convolutional network with the help of decoder this network is used for up-sampling of the input feature maps to restore and reconstruct the input size. Specifically, the decoder uses the polling indices that are calculated at the maximum polling phase of the respective encoder to perform nonlinear up-sampling. One or more conventional layers with a ReLU for non-linearity and with batch normalization composed the architecture of SegNet (in Figure 3) that consist of a sequence of encoder (non-liner processing layer). The encoder learns the representative features in the input image, and a compatible set of decoder layers that is followed by a pixel classifier. In the encoding sequence, the encoding done with the pooling process is up-sampled in the decoder with the help of max pooling.

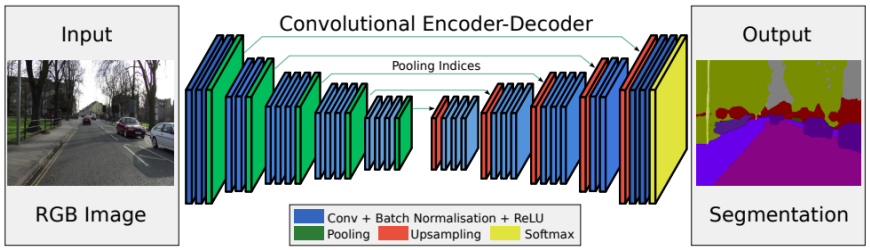

Fig. 3. SegNet architecture [10]

UNet. In [11], The encoder is primarily made up of a contracting path (the encoder), which captures the context of the image and an expandable symmetric path (de-

coder) to enable accurate location. Compared to SegNet, UNet use skip connections and concatenate features from low level to high layers to preserve spatial information.

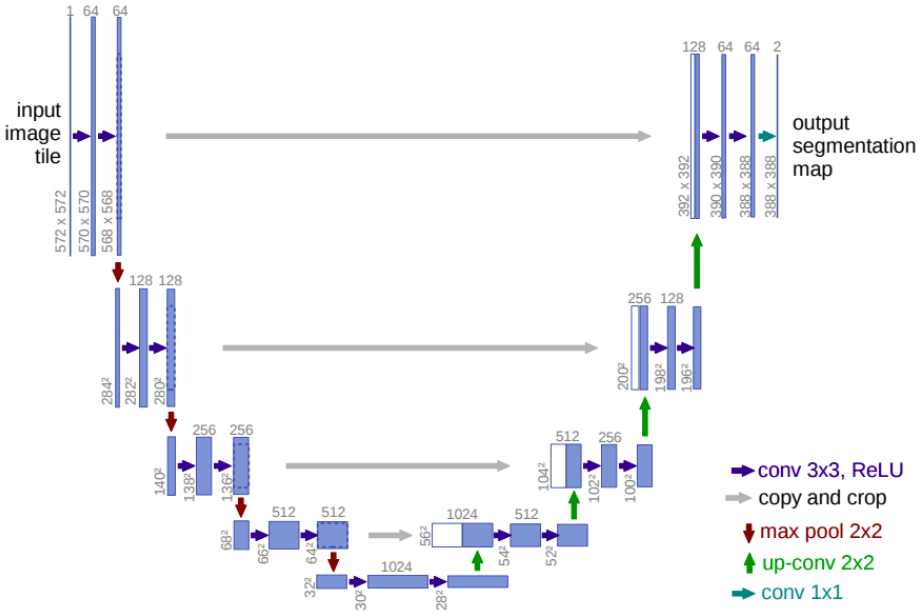

Fig. 4. UNet architecture [11]

Deeplab. [12] An architecture is presented by this network for learning of multi-scale contextual features and controlling of signal decimation. The model combines the advantages of dilated convolutions combined with feature pyramid pooling for multi-scale. From the last few max-pooling layers of DCNNs they remove the down-sampling operator and in the feature maps the resulting is computed at higher rate of sampling. Fully connected CRF is used for capturing the fine details. The Conditional Random Field's capabilities include smoothing terms that maximize the agreement of labels b/w similar pixels and can incorporate more broad terms that exemplify the contextual relationship b/w object classes.

Atrous convolution $y[i]$ is defined as:

$$y[i] = \sum_{k=1}^{K} x[i + r.k]w[k] \quad (1)$$

In the above equation One-dimensional signal is denoted by x[i], Length is denoted by K, Filter is denoted by w[k] And stride rate is denoted by r.

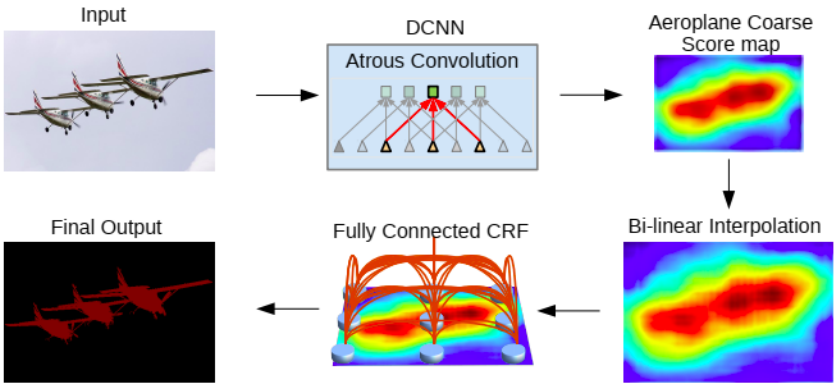

Fig. 5. Deeplab V1 architecture [12]

Deeplab V2. to represent the object in multiple scales, they offer resized DCNN versions of the same image and then combine score maps or features. Deeplab V2 use the ASPP. The idea is to apply multiple atrous variables at a different rate of sampling to the input feature map and grouped together. Atrous Spatial Pyramid Pooling helps to calculate the scales of various objects that can improve accuracy.

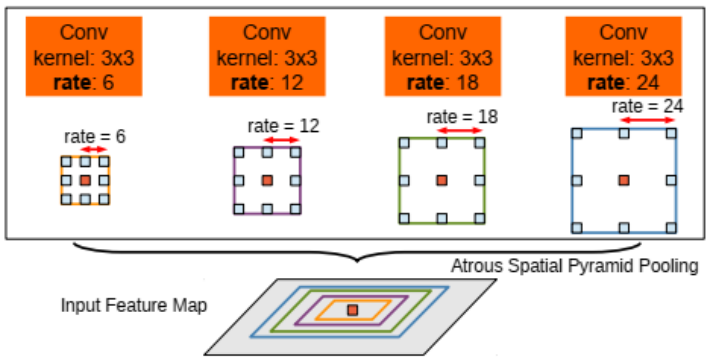

Fig. 6. Atrous Spatial Pyramid Pooling (ASPP) [12]

Deeplab V3. [13] the proposed ASPP proposed in Deeplab V2 makes the network able to encode the information of multiple scales by investigating incoming features through filters or atrous convolution functions by multiple effective fields and multiple rates. The next challenge was to gradually capture the boundaries of the object by retrieving local information. Deeplabv3 architecture adopts a novel encoder-decoder to solve this problem with Atrous Separative Convolution. The encoder-decoder model is capable of achieving sharp object parameters. Conventional encoder-decoder networks have been used successfully in many computers detection tasks, including human position measurement, object detection, and semantic segmentation.

In general, encoder and decoder networks consist of a module that gradually reduces the map element and captures high semantic information. - A decoder module

that slowly returns location information. In addition to the above encoded network, we also use highly sophisticated split-key modifications to enhance computer performance. This is achieved by making the standard conversion into a deep dynamic followed by a dynamic (i.e., 1 × 1 convolution). In particular, a highly intelligent change creates a spatial change independently of each input channel, while a clever transformation is used to integrate the output from the intellectual transformation.

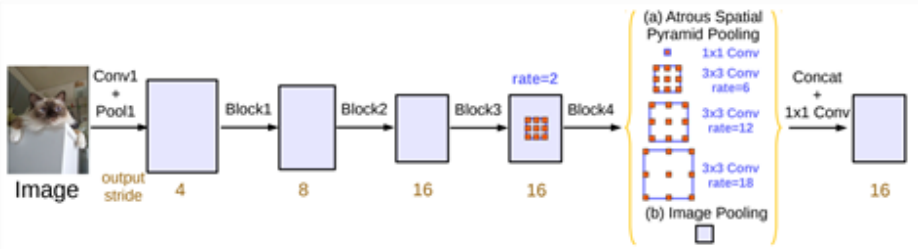

Fig. 7. Deeplab V3 architecture [13]

In particular, DeepLabv3 + enhances DeepLabv3 by adding a simple but efficient decoder module (Figure 7) to improve segmentation results, especially pyramid features and dilated convolutions is used along the object boundaries. Compressed prediction is obtained only by up-sampling the computing pixel-wise loss and the output of the last layer of convolution. Atrous convolution is applied by Deeplab to upsample.

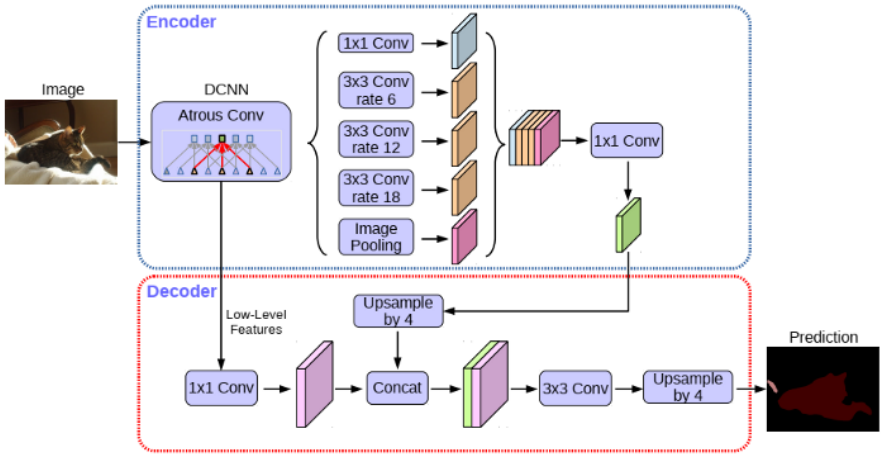

Fig. 8. Deeplab V3+ architecture [13]

The Table 1 presents a non-exhaustive list of the most well-known models used for supervised segmentation.

Table 1. Most known semantic segmentation deep models

Method	Year	Paper	Architecture
FCN	2014	Fully Convolutional Networks for Semantic Segmentation [1]	Transposed Layer
SegNet	2015	SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation [10]	Encoder/Decoder
UNet	2015	U-Net: Convolutional Networks for Biomedical Image Segmentation [11]	Encoder/Decoder
Mask RCNN	2017	Mask R-CNN [14]	Transposed Layer
PSPNet	2017	Pyramid Scene Parsing Network [15]	Transposed Layer
Deeplab	2018	DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs [12]	Atrous Convolution

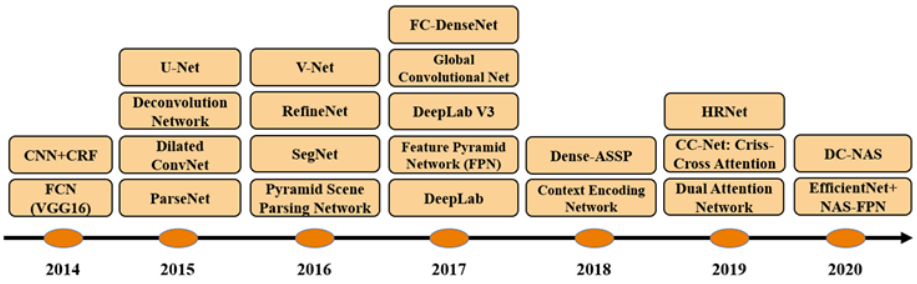

Fig. 9. The timeline of DL-based segmentation algorithms from 2014 to 2020

Networks backbones. many based CNN deep networks have been adapted to be used as backbones for semantic segmentation models, the backbone is the first part for any semantic network that is used to learn characteristics features of input images. The VGG network was proposed by [16] the VGG at Oxford University. It proposed according to layer number with different version with different versions, such as VGG-16, VGG-13, and VGG-19, the last full connected layer is removed and replaced by a decoder network for semantic segmentation task. ReNet [17] they replace convolution+pooling layers with multi-direction RNNs that sweep vertically and horizontally in both directions of the image. ResNet [18] achieves the better performance by successfully enabling a much deeper network in various vision tasks. In the modeling of the residual blocks the main contribution lies, with the help of this deep network structure can be trained easily, many versions have been proposed ResNet-50, ResNet100 and ResNet152. DenseNet [19] connects all the layer to one another. Advantages of DenseNet comes under the following aspects: 1) more reuse of features, 2) less parameters, and 3) a better training process that eliminates the problem of missing inclinations and model degradation. The aim of ResNeXt. Is to enhancing performance of the network while maintaining the complexity of the network, ResNeXt [20] have few hyper-parameters to set, and it is featured in its homogeneous, and multi-branch architecture. In this context, for balancing computational cost and accuracy, various lightweight networks designed with the help of MobileNet. Mo-

bileNetV1 [21] introduces in-depth flexibility, which gains significant improvements in efficiency. Faced with the limit of MobileNetV1, MobileNetV2 [22] is based on a distorted residual structure. MobileNetV3 [23] achieves the best performance with the smallest parameters by combining the attention method.

3 Weakly supervised methods: State of art

Weakly supervised learning methods are a set of models which attempt to build predictive models by learning with weak supervision. It consists on an approach to inject domain expertise into models or encompass a variety of training annotations less informative than the pixel level in order of decreasing informativeness. The most known techniques are using two sources of information priors and hints. Priors is what we believe to be true independent to any particular image sample or annotation, it what we know about the problem before even started looking at the data we have collected, while the hints are the indirect supervision received for each image on which we have some annotations. Priors are coming either implicitly or explicitly in the system, papers sometimes don't mention it but it will be coded in hyper-parameters or in some of aspect of how the model architecture operates. WSL methods for semantic segmentation have attracted a lot of interest due to the lack of fully annotated data for segmentation.

The Table 2 lists a no exhaustive priors and hints.

Table 2. Most common priors and hints

Prior	Hint
Size [24]	Image labels
Shape [25]	Image captions [25], [26]
Location	Transfer across images
Number of instances	Video labels
Contrast (boundaries, saliency) [25]–[29]	Click inside object
Class distribution [30]	Size from center click
Motion	Object bounding box
Similarity across images [31]	Objects extreme points
Similarity with external images	Scribbles [32]
	Eye Gaz
	Localized narrative

3.1 Challenges

The most common challenges that researchers encounter in weak supervision can be resumed to: i). The diverse appearance of object in same class when same object have different sizes and in different, ii). It not always obvious to estimate what is the full extent of the instance even if you use some kind of boundary estimation, iii). Another difficulty is that the semantics of what is being annotated it not so obvious,

especially when we are using scribbles as prior, iv). Also, the co-occurrence of elements of object to annotate when an object have many parts in the image.

In the technical side weakly, supervised paper must answer the bellow questions:

- Which priors are being used? this requires data and problem understanding
- How are these encoded?
- Which information source is used?
- Why was not that source exploited before?

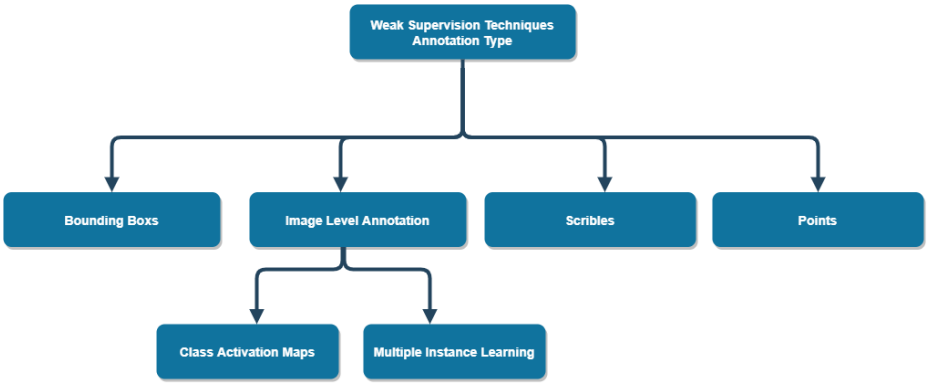

Fig. 10. Popular weak supervision methods

3.2 Image-level labels to pixel-level labels

Images labels only represent categories of objects present in an image in the form of words that are easy to collect and prepare manually or by automatic annotation techniques. However, they provide only a relatively low degree of supervision to a learning system. The ground truth's low dimension is a problem for modeling the semantic segmentation task in this class of methods. Indeed, the model input would be in the case of a single category of the form 1x1 and have to produce a probability map of the size of the images to segment. The following are some proposed techniques to address this problem.

Pseudo supervision (techniques based on class activation maps). This approach known as CAM (Class Activation Maps) techniques [33], [34]. The idea behind this approach is that even if the model is trained on image-level labels the CNNs can remarkably be able to identify objects in these approaches (Figure 11) the objective is to indicates the discriminative image regions to create new masks for segmentation training. It consists of training a classifier and global pooling layer or a spatial average pooling layer, an alternative of connected layer at the last convolutional layer, that behaves as a structural regularizer and prevent overfitting while training. Then one more convolution which will provide the per class coarse at the very end, it's latent spatial distribution of the things that interest for the class that you want to classify.

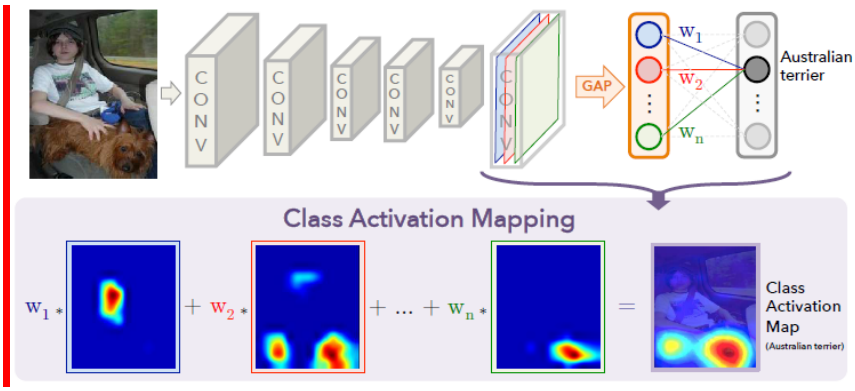

Fig. 11. Class Activation Mapping: The estimated class score is mapped back to the prior convolutional layer to produce the class activation maps (CAMs). The output of the CAM is the class-specific areas that is discriminative [33]

Last step is to threshold the class maps, high scores will represent the foreground areas and the low scores represents the background areas (Figure 12).

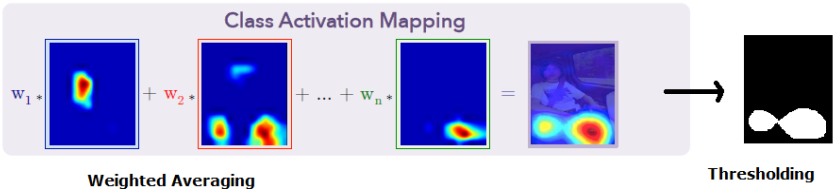

Fig. 12. Thresholding last scores to generate masks

In practice, these techniques have two steps, the first step consist of training a classification network (the backbone) to produce Class Activation Map (CAM) then training a segmentation network.

However, the CAMs allow to focus on discriminative area in the image where relevant feature is located, then result masks are not sharp and blobby, but it's not guarantee that will have a bias towards covering the full extent of the object, that can be improved by some techniques. To enhance the generated mask segmentation, [25] propose SEC methods that use an objective function sum of three loss functions seeding, expansion and constrain-to-boundary loss. L_{seed} the first term, gives the hints of localization to the network, the following term, L_{expand}, prohibit the network for the prediction of segmentation masks with very small or incorrect objects, and the last function $L_{constrain}$, helps in segmentation that acknowledge the color and spatial structure of the images, CRF is used in test step to leverage local information and global context. Another way is to use saliency information, [27], [28] propose combine saliency masks with image level tags to train a semantic network and to generate a pixel-wise labelling of object classes at test time. Image saliency is an image that highlights the region on which people's eyes focus first and has multiple connotations, it can refer to a spatial probability map, a probability map, or a binary mask. Other

methods are based on the idea that CAMs should be the same if the input image has small translations or deformations, To enforce the equivariance [26] propose SEAM method it consist of the integration of equivariant regularization and pixel correlation module (PCM) to keep consistent over affine transformation but also well fit the object contour. CAMs specially focus only on discriminative areas of the object of interest, to force the classifier to look more on neighbors areas of the object, one of the techniques used consists of hide some part in the input images in training stage, [30] proposed the division of every training image into grid of patches. Every patch has then given some random probability and feed as an input to a CNN system. During various epochs the hidden patches vary randomly. A complete image without any hidden patches is provided as input during testing, to the trained network which gives a classified annotation and object localization heatmap. In the same way [35] suggest an extra dropout layer, to enhance the model, before the first 1x1 convolutional layer. This dropout layer is added to improve the generalization behavior of the model. It prevents activations from becoming deeply correlated which leads to over-training of the model. Instead of doing random dropout [36] proposed a method to train the classification network first with the existing processed image and then a classification activation method is used to generate the class-specific response heatmap. To bring out the discriminative region a hard thresholding on the heatmap was applied. Then erases the discriminative region from the input image and fed it again into the classification network for learning to localize a new discriminative region then repeat the process until the discriminative region dropout significantly. The extracted areas from several steps together represent the predicted object regions as an output, which if later utilized for training the segmentation network. For the optimization of CAM, the most popular approaches just assess the problem that CAM can only trigger the sparse and discriminative regions for each class. Though, the classification network is weak in capturing related contextual information because the loss function of the classification task is image-level supervision, which leads to a new problem that several wrongly classified regions are activated in CAM. [37] implement two-stage training techniques. First to optimize the CAM that is formed by the multi-label classification network to produce pseudo ground truth. Second to train a fully supervised conventional semantic segmentation network through pseudo ground truth. [37] propose to use superpixels in mining semantic affinities between pixels, they use the assumption that pixels belonging to the same superpixel often have the same class label. [38] suggested an Atrous Convolutional Feature Network (ACFN) that produce dense object attention maps. The approach is to enhance the context representation of image classification CNNs. More precisely, cascaded atrous convolutions were applied in the middle layers to maintain necessary spatial features. Pyramidal atrous convolutions were also used in the end of convolutional layers to support multi-scale context information to extract the object attention maps. [29] put forward a simple to complex approach. The method is that the initial DCNN and the improved DCNN gradually and increasingly drive a simple feature map to a better pixel-level supervision. The features are then feet to the segmentation network. [40] introduced a proposal aggregation block and converted the mask generating into the task of regional proposal classification, where they applied the idea to aggregate the classified proposals. [41] introduced an iterative bottom-up and top-down framework. The name of the frame-

work is MCOF, which iteratively exploits the common object characteristics from the initial salient region. AffinityNet is proposed by [42], and this approach is dependent on the semantic affinity to propagate the local response generated by the classification network. They use CNN and execute a random walk from the seed and propagates the class activations in a asemantic affinity graph. This provides a pseudo ground-truth for training a FCN. CoDNet model and MR-CAM algorithm was proposed by [31]. In CODNet the input is a pair of images and then it extracts the common semantic and inter-sample similarity. In the target image features, for each location, from similar areas in the reference images are extracted and added to the original features. Through the inference, the MRCAM approach enhance segmentation masks by using multiple target-reference image pairs. [43] applies this two-step iterative method in Expectation-Maximization (EM) framework. In this framework the pixel-level labeling is considered as latent variables to be taken approximate from known image-level annotations (E-step). Using stochastic gradient descent and a probability distribution, the method then updates the neural network parameters and combines a pixel distribution and an adaptive bias (M-step). As [44] described that when it comes to take advantage of the full power of weak labels, the EM-Adapt method has limitations. This problem is generally non-convex, according to the authors, and needs Lagrangian dual optimizations which requires high computations. This proposed approach finds a way around the dual Lagrangian optimization as they integrate the constraint at network output level. Alternatively, the authors cast the segmentation task onto a constraint optimization problem in which the CNN network parameters are found given particular constrain Q with respect to the weak annotations. FickleNet [39], used a center-fixed spatial dropout in subsequent layers (by dropping out the non-center pixels in each convolutional layer) and trained a CNN at the image level. To create a threshold pseudo ground-truth for training an FCN the authors then run Grad-CAM several times. To basically, propagate class activations from high-confidence areas to neighboring areas with common visual appearance is another solution. DSRG [40] proposed a trained CNN and used region-growing on the generated CAMs to create a pseudo ground-truth and train an FCN. IRNet [41] introduced a similar approach but pursues to segment the instances individually by performing the random walk from low-displacement field centroids in the CAM seeds up until the class borders the used it as the pseudo ground-truths for training an FCN.

Multiple instances learning techniques. To handle weakly annotated data in the form of sequence level ground truth Multiple Instance Learning (MIL) is used [46]. In MIL there are bags which are the training examples organized in sets. An entire bag has a label, opposedly to the instances themselves based on an assumption [42]. In semantic segmentation the method learn to predict classes present in an image (known as a "bag") given ground-truth image-level labels and then, given the knowledge that at least a single pixel of every class is present, allocating pixels (known as "words") to each predicted class. This kind of approach often refers to train a CNN with image-level loss and inferring the image locations liable for each class prediction. Generally, MIL has a learning model that embeds individual instance of a bag into a laten space. After embeddings, the collection "usually of fixed sized" of instance latent vectors are sent into an aggregating function. This function outputs the predicted bag probability by using various principles, such as support vector machine, max-pooling or even

attention based neural networks. Two ways exists for interpreting multiple instance learning. MIL for classifying bags or slides and MIL for training an example classifier model, clear to bag segmentation. Particularly, to first train an instance model, researchers used max-pooling MIL techniques and its relaxed formulation and then investigate different ways to merge instance predictions into a slide prediction. [48] proposed a system in which the authors implement MIL with FCN jointly through a building the multi-class MIL loss. First, they obtained a 1 x 1 x C global class-aware vector by pixel intelligently extracting the highest value along the C direction. After this, with the obtained vectors the MIL loss is built by using the cross-entropy function. [49] proposed a more enhanced smooth form called log-sum-exp (LSE) which is a different approach from [48] that used the max function to extract the class-aware vectors. Two more priors were taken into consideration [49] in the test time, the tag-level prior and smoothing priors to produce more fine-grained results. [44] proposed many different extra constraints. For example, suppression constraint, foreground constraint, and size constraint. These constrained was introduced for the training of neural network, separately from the tag information. Since the forementioned methods successfully used image tags to understand semantic segmentation, their accuracy is still much lower than the performance of fully supervised methods. [49] presumed object segmentation by leveraging only object class information and considering only minimal priors on the object segmentation task. [50], [51] trains a CNN at the image level and to achieve the course class activation maps at the first and intermediate convolutional layers the authors uses guided back-propagation (GBP), then minus the maps from each other and takes the average of the maps across various scales and layers, followed by CRF post-processing. In [52] the author trained an FCN with a foreground and background mask that is generated by CRF on the scaled average of middle convolutional layers' features with the cross-entropy loss amidst the image-level labeling and the LSE pool of foreground and background masked features, CRF post-processing is applied at the test time. In [53] approach an FCN was trained with conv5 features that are further fed into a WSL transfer network. After this a class-wise average pooling and weighted spatial average of top and lowest activating activations were applied at the test time. This infers the maximum scoring class per position and pos-processing with CRF. [43] combine saliency and attention maps obtain reliable cues capable of significantly boosting the performance.

Table 3. Image Level based techniques results in Pascal VOC 2012 dataset

Annotation type : Image-level			
Technique	***Mehtod / Contributions***	***Publish***	***mIoU%***
CAM	Pathak et al. [44] Constraints output loss function	2015	45.1
	Pinheiro et al. [45] Agregation layer	2015	40.6
	Papandreou et al. [46] EM methods	2015	
	Tompson et al. [35] Spatial dropout layer	2015	

	Zhou et al. [33] Spatial average pooling layer	2016	41.0
	Kolesnikov et al. [25] SEC (seeding loss, expansion loss and constrain-to-boundary loss)	2016	50.7
	Qi et al. [47] Proposal aggregation and selection modules	2016	50.41
	Wei et al. [29] Simple-to-complex method	2017	44.9
	Oh et al. [27] Use of saliency masks	2017	56.9
	Chaudhry et al. [43] Saliency and attention maps	2017	60.8
	Singh et al. [30] Hide and seek using patches	2018	57.1
	Wei et al. [36] Mine regions by erasing discriminative regions	2018	55.7
	Wang et al. [48] MCOF	2018	56.2
	Ahn et al. [49] AffinityNet	2018	61.7
	Ahn et al. [41] IRNet	2019	63.5
	Wang et al. [26] SEAM method	2020	65.7
	Chang et al. [37] Superpixels correlation affinities	2020	64.8
	Xu et al. [28] Saliency masks	2021	69.0
	Wan et al. [31] Co-attention : CODNet model and MR-CAM	2021	64.5
	Xu et al. [38] Atrous Convolutional Feature Network	2021	66.6
	Chong et al. [50] Two stage training to optimize CAMs	2021	66.8
MIL	Pathak et al [44] MIL–FCN with Multi-class MIL loss	2015	25.66
	Pathak et al. [51] Extra constraints for suppression foreground constraint, and size constraint.	2015	45.1
	Pinheiro et al. [45] Log-sum-exp (LSE)	2015	40.6
	Saleh et al [52] FCN with a foreground/background mask	2016	46.6
	Durand et al [53] WSL transfer network and weighted spatial average	2017	53.4
	Shimoda et al [54] GBP (guided back-propagation)	2020	51.3

3.3 Methods based on bounding-box-level supervision

Bounding box annotation provides the completed location of a whole object, as well as its semantic tag. The main idea is to obtain pseudo masks by adopting an unsupervised or semi-supervised segmentation algo from the bounding box annotation.

[55] applied CRF [56] withing bounding boxes and extracted the pseudo masks. The expectation maximization (EM) algorithm is then used to enhance the pseudo masks produced. [57] implements the region proposal methods [58], [59] to create the candidate pseudo mask. The segmentation network is trained based on this in an iterative approach. Precisely, under the supervision of candidate masks, the segmentation network is first trained and then choose the better masks for the next training repetition. [60] utilized a Simple Does It (SDI) which is a repetitive training method to slowly enhance generated label estimates. But SDI uses a GrabCut like algorithm for the very first lable estimate creation, BoxSup uses an unsupervised area proposal approach like Multiscale Combinatorial Grouping (MCG) [59]. Furthermore, BoxSup changes the training process in order to remove the noise the middle layer outputs. SDI leaves the training algorithm unchanged and concentrates on externally removing noisey input labels through utilizing prior knowledge. [60] obtained the pseudo mask by applying Grabcut to bounding boxes. Not only using CRF [62] introduced the box-driven class wise masking mode (BCM), with the filling rate guided adaptive loss (FR0Loss), trying to get rid of the incorrectly annotated regions in the pseudo mask. [55] introduced a BB-UNet ("Bounding Box U-Net"), which is a deep learning model that incorporates location and also shape prior onto model training. U-Net is the inspiration for the proposed method, and it integrates priors through a novel convolutional layer proposed at the level of skip connections. From the performance of a trained object detector, [64] uses higher level information, by striving for the smallest locations of the image from which the object detector then creates almost the equal result because it does from the entire image. These regions form a "bounding-box attribution map" (BBAM), which recognizes the target object in its bounding box and thus acts as pseudo ground-truth.

Table 4. Bounding boxs based techniques results in Pascal VOC 2012 dataset

Mehtod	Contributions	Publish	mIoU%
Rother et al. [56]	GrabCut	2004	
Papandreou et al. [46]	CRF + Bounding Boxes + EM	2015	62.2
Dai et al. [57]	Region proposal methods to generate pseudo masks	2015	63.8
Khoreva et al. [58]	Simple Does It (SDI)	2017	67.5
Song et al. [59]	Box-driven classwise masking model	2019	67.5
Rosana et al. [55]	BB-UNet	2020	68.2
Lee et al. [60]	Bounding-box attribution map (BBAM)	2021	73.7

3.4 Methods based on scribble-level supervision

Scribble-supervised semantic segmentation aims to produce dense predictions given only sparse scribbles. Despite the feasibility of the tag-based methods, their accu-

racy is still not satisfying, due to the very limited tag-level information. Scribbles compromise between image tags and pixelwise annotations. Compared to image tags, scribbles use limited pixels to provide location information. Compared to pixelwise annotations, scribbles cost much less manual labeling efforts. To some extent, scribbles can be seen as the combination of image tags and a set of fully annotated pixels. In this part, we mainly review the methods based on two kinds of scribbles, i.e., point scribbles and line scribbles.

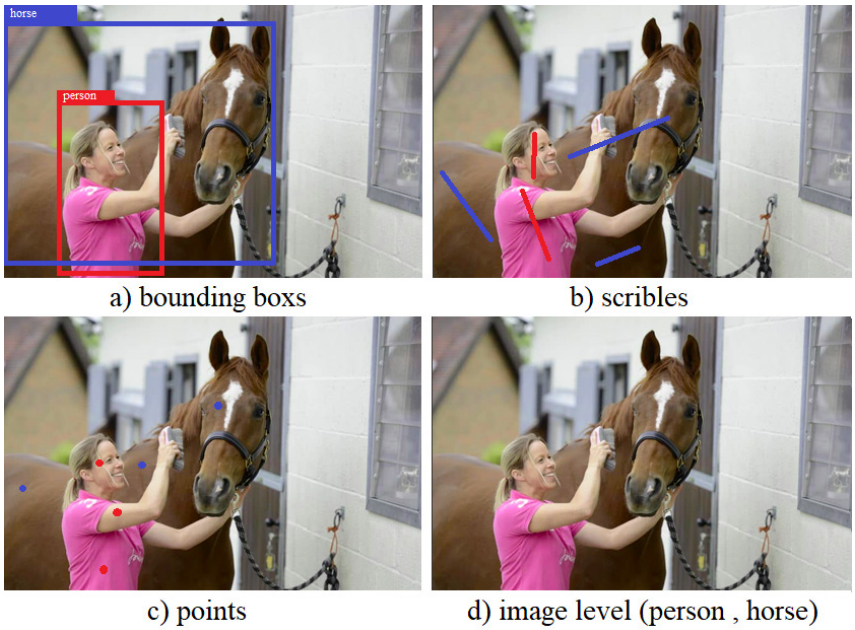

Fig. 13. Different kind of weak labels. In order of decreasing informativeness: (a) Bounding box, (b) Scribble, (c) Point and (d) Image-level label

Points priors. To enhance the behavior and capability of segmentation algorithm, point supervision is the key idea. The collection is simple when labeling a new dataset. A novel loss function was proposed by [66] to lead the network training, that has two components, "the loss for the tag-level inference" and "the loss calculated from the point scribbles". To enhance it for the production of more precise results the authors further include a generic objectness prior [67] in the loss function, i.e., what is the probability of a pixel belong to an object. [61] use point level annotations to achieve semantic segmentation; they propose the use of extreme points rather than more random points. [69] used and approach to generate the semantic segmentation from images given some point-level labeling. The method consists of annotating point in the training of CNN to produce enhanced localization and class activation maps. Another CNN is used to predict the semantic affinities to propagate rough class annotations and create pseudo semantic segmentation labels. [70] supports the semantic relationship between the labeled points by fostering the feature representations of the intra and inter category points to maintain stable and coherent. For example, points

within the same classified group should have more related feature meanings compared to those from other categories which leads to a simple distance metric loss, which work together with the point-wise cross-entropy loss to enhance the deep neural nets.

Scribbles priors. Most of the line scribbles-based methods consists of two parts. One part is scribble propagation and the other one is segmentation network. The scribble propagation part propagates the scribbles to the unannotated pixels; therefore this produce full pixel-level labels automatically, which are utilized further for the segmentation network training. The scribble propagation block is the main issue same as the pseudo supervised tag-level methods. ScribbleSup [62] is the first that proposed the technique of using deep learning into the scribble-supervised segmentation. It first produces an entire annotation map by utilizing the weakly annotated scribbles and a CRF model. Then the next phase is to implement the optimization of neural nets and CRF energy function alternately to improve the segmentation results. RAWKS [63] approach is to embed a deep segmentation network and a label-propagator that is learnable to gradually and increasingly update the segmentation network and generate dense labels. [62] utilized the graph model to generate the scribble to the unannotated pixels based on the constraints of spatial, visual, and semantic context characteristics. [72] proposed a graph convolutional network (GraphNet) model. First, the authors embedded the scribbles into the graph, then these embedded scribbles are fed into the network to create the pseudo mask. [73] proposed the random walk [74] to achieve the label propagation.

Table 5. Summarization of priors techniques

Annotation type : Scribles			
Technique	***Mehtod / Contributions***	***Publish***	***mIoU%***
Scribble	Krähenbühl et al. [63] RAWKS network	2012	30.2
	Lin et al. [62] ScribbleSup	2016	42.0
	Lin et al. [62] CRF energy function	2016	63.1
	Vernaza et al. [64] Random-walk label propagation	2017	61.1
	Pu et al. [65] GraphNet	2018	68.9
Point	Bearman et al. [66] Generic objectness prior	2016	46.1
	Maninis et al [61] Extreme points	2018	73.2
	Qian et al. [67] Point-based Distance Metric Learning (PDML)	2019	30
	McEver et al. [68] PCAM	2020	70.5

4 Evaluation metrics for semantic segmentation

Our next focus is on the assessment of metrics from two perspectives, that are accuracy and efficiency. For every perspective, the frequently utilized metrics are presented in the following. De denote TP (True Positive) the number of pixels that belong to foreground and was classified correctly, TN (True Negative) the number of pixels that belong to background and was classified correctly, FP (False Positive) the pixels that belong to foreground and classified as background, FN (False Negative) the pixels that belong to background and was classified as foreground.

The term **"Accuracy"** is the simplest to learn and understand conceptually. In the image accuracy is the percent of pixels that are categorized properly.

$$ACC = \frac{TP + TN}{TP + TN + Fp + FN} \tag{2}$$

While it is easy to understand, it is not the best metric to evaluate objectively the model performance, this is due to imbalanced data in images, in the most cases background pixels are considerably bigger than foreground pixels, accuracy can reach high value even when the model didn't correctly classify any foreground pixel.

Intersection-Over-Union (IoU, Jaccard Index). The Intersection-Over-Union (IoU) is also called the Jaccard Index. It is one of the most applied metrics in semantic segmentation. The IoU is a very simple metric that is exceedingly efficient, it is a number from 0 to 1 that specifies the amount of overlap between the predicted and ground truth bounding box.

$$IoU = \frac{|Prediction \cap GroundTruth|}{|Prediction \cup GroundTruth|} = \frac{TP}{TP + FP + FN} \tag{3}$$

For binary (two classes) or multiple-class segmentation, the represent the IoU of the image is determined by taking the IoU of each class and averaging them.

Dice Coefficient (F1 Score). Simply put the Dice Coefficient is 2 * the Area of Overlap divided by the total number of pixels in both images.

$$Dice = \frac{2.|Prediction \cap GroundTruth|}{|Prediction \cup GroundTruth|} = \frac{2.TP}{2.TP+FP+FN} = \frac{2.IoU}{IoU + 1} \tag{4}$$

IoU is a similar method to Dice coefficient. They have a positive correlation, which means that if one result, in a comparison of two models, that model C is better than model D at segmentation level of image, the other will also result the same. Similar to IoU, both of them range from 0 to 1. 1 shows the greatest similarity among the predicted and truth. A temptation conclusion that the two metrices are equivalent so the choice among them is optional, but now so quick. Problem arises while taking an average score over a set assumption. The gap occurs when calculating how much worse classifier D is than C for a particular case. Generally, the IoU metric has the tendency to penalize single instances of bad categorization more than the F-Score quantitatively even when both of them that a particular instance is bad. Likewise, to how L2 can penalize the largest errors more than L1, the IoU metric has the tendency to have a "squaring" effect on the mistakes compared to the F score. So, the F score

has tended to calculate something nearer to the average performance, while the IoU score measures something nearer to the worst-case performance.

5 Loss functions for semantic segmentation

Loss functions have an important role in machine learning models, and especially deep learning use stochastic gradient descent to optimize and learn the objective. By minimizing the loss function we evaluate the model against the learned parameters and defines how is good a model. Choosing the adequate loss function for a specific task is primordial to achieve the best prediction performance. In semantic segmentation we need to know if the loss function used is able to cover the edge cases.

In the next, we denote, p_i and g_i represent pairs of corresponding pixel values of prediction and ground truth, respectively ground truth and predicted segmentation, respectively.

5.1 Dice loss

The Dice coefficient is widely used metric to calculate the similarity between two images. And have been adapted to Dice loss that aims to maximize the overlap between two sets.

$$L_{\text{Dice}} = \frac{2\sum_i^N p_i g_i + 1}{\sum_i^N p_i^2 + \sum_i^N p_i^2 + 1} \tag{5}$$

Dice loss, is a better alternative to Cross Entropy Loss for boundary detection, while Cross Entropy Loss considers locally rather than considering it globally, which is not enough for image level prediction

The 1 is added in numerator and denominator to ensure that the function is not undefined in edge case scenarios such as when $p_i = g_i = 0$

5.2 Focal loss

This loss is an improvement to the binary cross-entropy. This loss function down-weights the contribution of easy examples using a modulating factor, this enables the model to focus more on learning hard examples to ensures that there is no class imbalance. The factor automatically down weights the contribution of easy examples at training time and focuses on the hard ones.

$$L_{FL} = -\frac{1}{N}\sum_{i=1}^{N} (1 - p_i)^{\gamma} g_i \log p_i \tag{6}$$

5.3 Intersection over Union (IoU)-balanced loss

The IoU-balanced classification loss aims at increasing the gradient of samples with high IoU and decreasing the gradient of samples with low IoU. This loss is similar to dice.

$$L_{IoU} = 1 - \frac{\sum_{i=1}^{N} g_i p_i}{\sum_{i=1}^{N} (g_i + p_i - g_i p_i)} \tag{7}$$

5.4 Boundary loss

Dice or cross-entropy are based on integrals over the segmentation regions. The boundary loss, which takes the form of a distance metric on the space of contours, not regions. To compute the distance Dist $(\partial G, \partial S)$ between two boundaries in a differentiable way, boundary loss uses integrals over the boundary instead of unbalanced integrals over regions to mitigate the difficulties of highly unbalanced segmentation.

$$L_{BD} = \int_{\Omega} \phi_G(p) s_\theta(p) dp \tag{8}$$

Where $\phi_G$is the level set representation of boundary: $\phi_G = -D_G(q)$ if q ∈ G, and $\phi_G = D_G(q)$ otherwise. $s_\theta(p)$ is network softmax probability outputs.

5.5 Weighted cross-entropy

In this loss positive examples are weighted by a certain coefficient to involve class imbalance.

Other loss functions used for semantic segmentation are summarized in Table 5.

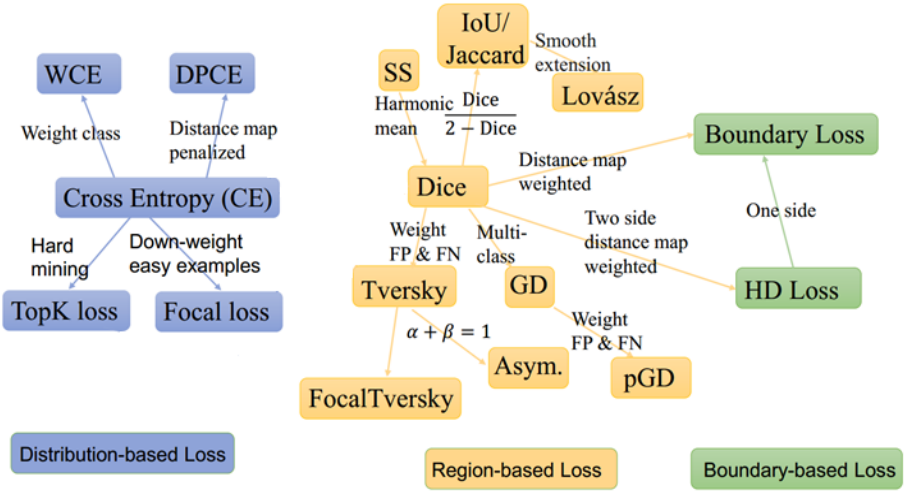

Fig. 14. Overview and relationship among the existing loss functions [69]

Table 6. The list of most knows loss functions for semantic segmentation

Type	Loss Function	Use cases
Distribution-based Loss	Binary Cross-Entropy	Works best in equal data distribution among classes scenarios
	Weighted Cross-Entropy	Give more weight to all positive examples
	Focal Loss	Works best with highly imbalanced dataset
Region-based Loss	Dice Loss	Inspired from Dice Coefficient, a metric to evaluate segmentation results
	IoU-balanced Loss	Avoid high unbalanced classes
Boundary-based Loss	Hausdorff Distance loss	Inspired by Hausdorff Distance metric used for evaluation of segmentation
	Shape aware loss	Variation of cross-entropy loss by adding a shape based coefficient used in cases of hard-to-segment boundaries.
	Boundary Loss	Takes the form of a distance metric on the space of contours, not regions. To compute the distance Dist $(\partial G, \partial S)$ between two between two boundaries in a differentiable way

6 Datasets

To analyze the performance of developed models in multiple image verities, many datasets are being considered in different studies for the project's experimental estimations. The list below is a non-exhaustive example of datasets for images semantic segmentation.

6.1 Atlas of Digital Pathology (ADP)

Huron is a leading company for development of the Tissue Scope mainly LE 1.2 used for microscopic examination. There is database of the microscopic examination images from the different body tissues. The said database is said to be known as The Atlas of Digital. There are bunch of images available for the experiment and training purpose with almost more than 25 structural types and shades. There are some functional shades also labeled for the segmentation.

6.2 PASCAL VOC2012

The most used dataset for semantic segmentation, object detection and classification models evaluation. The 2012 release of the PASCAL VOC challenge dataset is a natural scene ("in the wild") images captured by many consumer cameras in the world. Each image is labelled with 20 foreground classes, with an added background class for segmentation. This dataset contains 20 object categories including vehicles, household, animals, and other: airplane, bicycle, … The **PASCAL VOC** dataset is split into three subsets: 1,464 images for training, 1,449 images for validation and a private testing set.

6.3 DeepGlobe Land Cover Classification

If we want to discuss about the satellite imagery the only best multilabel choice, we have is the DeepGlobe Land Cover Classification dataset. It is the collection of the intense rich imagery from the across the Globe. The Land cover consists of more than 5 classes with one no Land cover one.

6.4 Common Objects in COntext—Coco dataset

The densest dataset in respect of image quantity is the MS COCO use for a massive detection and segmentation. The most prominent about this dataset is the quantity of the images which touches the 325K+ images shades. There is an explanation for the below type of modeling.

1. Object Detection: There are more than 70 objects for the Bounding-Box as well as Segmentation for each instance.
2. Captioning: Images have the description in the natural language to easily communicate the main idea about the image.
3. Keypoints Detection: There are more than 200K images and instances for each person including label for each with Keypoints. Grass, Sky, and other such things have more than 75 categories for the full segmentation.
4. Dense Pose: Dense pose have more than 35000 images in the dataset.

6.5 The Cityscapes dataset

If we want to have the test and training or even implementation approach for the pixel level and instance level segment labeling, then Cityscapes is a leading dataset best suited the purpose. The said dataset consists of more than 45 cities cameras videos. More than 4500 ultimate quality pixels with more than 18000 extra imageries with week base labeling make it a prominent dataset for training and testing.

6.6 The Cambridge-driving labeled video database—CamVid

The first collection of videos with object class semantic labels, complete with metadata. The database provides ground truth labels that associate each pixel with one of 32 semantic classes. The database addresses the need for experimental data to quantitatively evaluate emerging algorithms.

6.7 CHAOS

The segmentation of spleen, kidneys, and liver from MRI and CT data is the challenging aim of the CHAOS. On the 11 of April 2019 CHAOS was the part of IEEE International Symposium on Biomedical Imaging (ISBI) that was held in ITALY. 40 different patient's CT Images was the part of first database. These patients were donors of liver, but they have no lesion, no tumor nor any other diseases, and have healthy liver. 120 DICOM datasets were part of 2nd database that contains T1-DUAL

(40 out phase, 40 in phase) and T2-SPIR (40 datasets). By the help of gradient combination and different RF pulse each of the following scanned on daily bases.

6.8 The Liver Tumor Segmentation Benchmark (LiTS)

On 2016 in the conjunction with IEEE (ISBI) the Liver Tumor Segmentation Benchmark was organized and on 2017 an international conference was held in MICCAI and LITS was part of that conference. LITS algorithms were applied at the set of 131 CT volumes that has different types of tumor contrast levels that contains varying amount of lesion and tissues size abnormalities.

Table 7. The list of most know dataset for semantic segmentation

Dataset	Link	Annotation Type	Images Nature
CHAOS	https://chaos.grand-challenge.org	Ground Truth	CT+MRI
CAMVID	http://mi.eng.cam.ac.uk/research/projects/VideoRec/CamVid/	Ground Truth	Natural
CITYSCAPES	https://www.cityscapes-dataset.com/	Coarse annotations	Natural
COCO	https://cocodataset.org/	Ground Truth	Natural
ADP	https://www.dsp.utoronto.ca/projects/ADP/	Patches	TIFF
PASCAL VOC12	http://host.robots.ox.ac.uk/pascal/VOC/	Ground Truth	Natural
DeepGlob	http://deepglobe.org/	Ground Truth	Satellite
LiTS	https://arxiv.org/pdf/1901.04056v1.pdf	Ground Truth	CT

7 Conclusion

Image segmentation has made significant advances in recent years. Recent work based largely on deep learning techniques which has resulted in groundbreaking improvements in the accuracy of the segmentations. In this paper, we briefly review the deep-learning-based semantic segmentation methods from a different perspective, which are divided according to the supervision level. Some widely used deep learning architectures are investigated and we especially focused on weakly supervised methods. For each reviewed method, we provide details on its contribution, publishing year and results. We also discuss the common challenges and several possible directions in this field. We conclude the number of studies in weakly supervised has soared and the degree of attention has increased significantly in recent years. This is because time-consuming and labor-intensive pixel-by-pixel annotations are no longer sufficient for today's development needs, and people need to use more economical and efficient research methods. However, it can be seen from the analysis of the experimental results that the current methods still have shortcomings, and there are still many aspects to be further studied. Finally, from our perspective, the study of weakly supervised learning is to pave the way for the ultimate realization of unsupervised learning while improving the efficiency of fully supervised learning. So far, research on unsupervised learning has not been interrupted, whether in the field of image seg-

mentation or in other image fields, or even in the field of natural language processing. Because completing tasks without any label is the ideal state for machine learning.

8 References

[1] J. Long, E. Shelhamer, et T. Darrell, (2015). Fully Convolutional Networks for Semantic Segmentation, *arXiv*. https://doi.org/10.1109/CVPR.2015.7298965

[2] S. Hao, Y. Zhou, et Y. Guo, (2020). A Brief Survey on Semantic Segmentation with Deep Learning, *Neurocomputing*. https://doi.org/10.1016/j.neucom.2019.11.118

[3] A. Khan, A. Sohail, U. Zahoora, et A. S. Qureshi, (2020). A Survey of the Recent Architectures of Deep Convolutional Neural Networks, *Artif Intell Rev*, vol. 53, nº 8, p. 5455-5516. https://doi.org/10.1007/s10462-020-09825-6

[4] F. Lateef et Y. Ruichek, (2019). Survey on semantic segmentation using deep learning techniques, *Neurocomputing*, vol. 338, p. 321-348. https://doi.org/10.1016/j.neucom.2019.02.003

[5] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, et F. E. Alsaadi, (2017). A survey of deep neural network architectures and their applications, *Neurocomputing*, vol. 234, p. 11-26. https://doi.org/10.1016/j.neucom.2016.12.038

[6] G. Litjens *et al.*, (2017). A survey on deep learning in medical image analysis, *Medical Image Analysis*, vol. 42, p. 60-88. https://doi.org/10.1016/j.media.2017.07.005

[7] M. Thoma, (2016). A Survey of Semantic Segmentation, *arXiv:1602.06541 [cs]*.

[8] M. Zhang, Y. Zhou, J. Zhao, Y. Man, B. Liu, et R. Yao, (2019). A survey of semi- and weakly supervised semantic segmentation of images, *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-019-09792-7

[9] F. Yi et I. Moon, (2012). Image segmentation: A survey of graph-cut methods, in *2012 International Conference on Systems and Informatics (ICSAI2012)*, Yantai, China, p. 1936-1941. https://doi.org/10.1109/ICSAI.2012.6223428

[10] V. Badrinarayanan, A. Kendall, et R. Cipolla, (2015). SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation, *arXiv:1511.00561 [cs]*.

[11] O. Ronneberger, P. Fischer, et T. Brox, (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation, *International Conference on Medical image computing and computer-assisted intervention*, p. 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

[12] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, et A. L. Yuille, (2018). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, nº 4, p. 834-848. https://doi.org/10.1109/TPAMI.2017.2699184

[13] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, et H. Adam, (2018). Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, *arXiv:1802.02611 [cs]*. https://doi.org/10.1007/978-3-030-01234-2_49

[14] K. He, G. Gkioxari, P. Dollár, et R. Girshick, (2017). Mask R-CNN, arXiv*:1703.06870 [cs]*. https://doi.org/10.1109/ICCV.2017.322

[15] H. Zhao, J. Shi, X. Qi, X. Wang, et J. Jia, (2017). Pyramid Scene Parsing Network, *arXiv:1612.01105 [cs]*. https://doi.org/10.1109/CVPR.2017.660

[16] K. Simonyan et A. Zisserman, (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition, *arXiv:1409.1556 [cs]*.

[17] F. Visin, K. Kastner, K. Cho, M. Matteucci, A. Courville, et Y. Bengio, (2015). ReNet: A Recurrent Neural Network Based Alternative to Convolutional Networks, *arXiv:1505.00393 [cs]*.

[18] K. He, X. Zhang, S. Ren, et J. Sun, (2015). Deep Residual Learning for Image Recognition, *arXiv:1512.03385 [cs]*. https://doi.org/10.1109/CVPR.2016.90
[19] P. Bilinski et V. Prisacariu, (2018). Dense Decoder Shortcut Connections for Single-Pass Semantic Segmentation, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, p. 6596-6605. https://doi.org/10.1109/CVPR.2018.00690
[20] S. Xie, R. Girshick, P. Dollár, Z. Tu, et K. He, (2017). Aggregated Residual Transformations for Deep Neural Networks, *arXiv:1611.05431 [cs]*. https://doi.org/10.1109/CVPR.2017.634
[21] A. G. Howard *et al.*, (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications, *arXiv:1704.04861 [cs]*.
[22] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, et L.-C. Chen, (2019). MobileNetV2: Inverted Residuals and Linear Bottlenecks, *arXiv:1801.04381 [cs]*. https://doi.org/10.1109/CVPR.2018.00474
[23] A. Howard *et al.*, (2019). Searching for MobileNetV3, *arXiv:1905.02244 [cs]*. https://doi.org/10.1109/ICCV.2019.00140
[24] D. Zhang, K. Song, J. Xu, H. Dong, et Y. Yan, (2022). An image-level weakly supervised segmentation method for No-service rail surface defect with size prior, *Mechanical Systems and Signal Processing*, vol. 165, p. 108334. https://doi.org/10.1016/j.ymssp.2021.108334
[25] A. Kolesnikov et C. H. Lampert, (2016). Seed, Expand and Constrain: Three Principles for Weakly-Supervised Image Segmentation, *arXiv:1603.06098 [cs]*. https://doi.org/10.1007/978-3-319-46493-0_42
[26] Y. Wang, J. Zhang, M. Kan, S. Shan, et X. Chen, (2020). Self-Supervised Equivariant Attention Mechanism for Weakly Supervised Semantic Segmentation, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, p. 12272-12281. https://doi.org/10.1109/CVPR42600.2020.01229
[27] S. J. Oh, R. Benenson, A. Khoreva, Z. Akata, M. Fritz, et B. Schiele, (2017). Exploiting Saliency for Object Segmentation from Image Level Labels, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, p. 5038-5047. https://doi.org/10.1109/CVPR.2017.535
[28] L. Xu, W. Ouyang, M. Bennamoun, F. Boussaid, F. Sohel, et D. Xu, (2021). Leveraging Auxiliary Tasks With Affinity Learning for Weakly Supervised Semantic Segmentation, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, p. 10. https://doi.org/10.1109/ICCV48922.2021.00690
[29] Y. Wei *et al.*, (2017). STC: A Simple to Complex Framework for Weakly-supervised Semantic Segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, nº 11, p. 2314-2320. https://doi.org/10.1109/TPAMI.2016.2636150
[30] K. K. Singh, H. Yu, A. Sarmasi, G. Pradeep, et Y. J. Lee, (2018). Hide-and-Seek: A Data Augmentation Technique for Weakly-Supervised Localization and Beyond, *arXiv:1811.02545 [cs]*.
[31] W. Wan, J. Chen, M.-H. Yang, et H. Ma, (2021). Co-attention dictionary network for weakly-supervised semantic segmentation, *Neurocomputing*, p. S0925231221017252. https://doi.org/10.1016/j.neucom.2021.11.046
[32] X. Liu *et al.*, (2022). Weakly Supervised Segmentation of COVID19 Infection with Scribble Annotation on CT Images, *Pattern Recognition*, vol. 122, p. 108341. https://doi.org/10.1016/j.patcog.2021.108341
[33] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, et A. Torralba, (2016). Learning Deep Features for Discriminative Localization, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, p. 2921-2929. https://doi.org/10.1109/CVPR.2016.319

[34] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, et D. Batra, (2020). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, *Int J Comput Vis*, vol. 128, nº 2, p. 336-359. https://doi.org/10.1007/s11263-019-01228-7

[35] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, et C. Bregler, (2015). Efficient object localization using Convolutional Networks, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, p. 648-656. https://doi.org/10.1109/CVPR.2015.7298664

[36] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, et S. Yan, (2018). Object Region Mining with Adversarial Erasing: A Simple Classification to Semantic Segmentation Approach, *arXiv:1703.08448 [cs]*. https://doi.org/10.1109/CVPR.2017.687

[37] Y.-T. Chang, Q. Wang, W.-C. Hung, R. Piramuthu, Y.-H. Tsai, et M.-H. Yang, (2020). Weakly-Supervised Semantic Segmentation via Sub-Category Exploration, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, p. 8988-8997. https://doi.org/10.1109/CVPR42600.2020.00901

[38] L. Xu, H. Xue, M. Bennamoun, F. Boussaid, et F. Sohel, (2021). Atrous convolutional feature network for weakly supervised semantic segmentation, *Neurocomputing*, vol. 421, p. 115-126. https://doi.org/10.1016/j.neucom.2020.09.045

[39] J. Lee, E. Kim, S. Lee, J. Lee, et S. Yoon, (2019). FickleNet: Weakly and Semi-supervised Semantic Image Segmentation using Stochastic Inference. arXiv, 2019. https://doi.org/10.1109/CVPR.2019.00541

[40] Z. Huang, X. Wang, J. Wang, W. Liu, et J. Wang, (2018). Weakly-Supervised Semantic Segmentation Network with Deep Seeded Region Growing, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, p. 7014-7023. https://doi.org/10.1109/CVPR.2018.00733

[41] J. Ahn, S. Cho, et S. Kwak, (2019). Weakly Supervised Learning of Instance Segmentation with Inter-pixel Relations, *arXiv:1904.05044 [cs]*. https://doi.org/10.1109/CVPR.2019.00231

[42] J. Foulds et E. Frank, (2010). A review of multi-instance learning assumptions, *The Knowledge Engineering Review*, vol. 25, nº 1, p. 1-25. https://doi.org/10.1017/S026988890999035X

[43] A. Chaudhry, P. K. Dokania, et P. H. S. Torr, (2017). Discovering Class-Specific Pixels for Weakly-Supervised Semantic Segmentation, *arXiv:1707.05821 [cs]*. https://doi.org/10.5244/C.31.20

[44] D. Pathak, P. Krähenbühl, et T. Darrell, (2015). Constrained Convolutional Neural Networks for Weakly Supervised Segmentation, *arXiv:1506.03648 [cs]*. https://doi.org/10.1109/ICCV.2015.209

[45] P. O. Pinheiro et R. Collobert, (2015). From Image-level to Pixel-level Labeling with Convolutional Networks, *arXiv:1411.6228 [cs]*. https://doi.org/10.1109/CVPR.2015.7298780

[46] G. Papandreou, L.-C. Chen, K. P. Murphy, et A. L. Yuille, (2015). Weakly-and Semi-Supervised Learning of a Deep Convolutional Network for Semantic Image Segmentation, in *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, p. 1742-1750. https://doi.org/10.1109/ICCV.2015.203

[47] X. Qi, Z. Liu, J. Shi, H. Zhao, et J. Jia, (2016). Augmented Feedback in Semantic Segmentation Under Image Level Supervision, in *Computer Vision – ECCV 2016*, vol. 9912, B. Leibe, J. Matas, N. Sebe, et M. Welling, Éd. Cham: Springer International Publishing, 2016, p. 90-105. https://doi.org/10.1007/978-3-319-46484-8_6

[48] X. Wang, S. You, X. Li, et H. Ma, (2018). Weakly-Supervised Semantic Segmentation by Iteratively Mining Common Object Features, *arXiv:1806.04659 [cs]*. https://doi.org/10.1109/CVPR.2018.00147

[49] J. Ahn et S. Kwak, (2018). Learning Pixel-level Semantic Affinity with Image-level Supervision for Weakly Supervised Semantic Segmentation, *arXiv:1803.10464 [cs]*. https://doi.org/10.1109/CVPR.2018.00523

[50] Y. Chong, X. Chen, Y. Tao, et S. Pan, (2021). Erase then grow: Generating correct class activation maps for weakly-supervised semantic segmentation, *Neurocomputing*, vol. 453, p. 97-108. https://doi.org/10.1016/j.neucom.2021.04.103

[51] D. Pathak, E. Shelhamer, J. Long, et T. Darrell, (2015). Fully Convolutional Multi-Class Multiple Instance Learning, *arXiv:1412.7144 [cs]*.

[52] F. Saleh, M. S. A. Akbarian, M. Salzmann, L. Petersson, S. Gould, et J. M. Alvarez, (2016). Built-in Foreground/Background Prior for Weakly-Supervised Semantic Segmentation, *arXiv:1609.00446 [cs]*. https://doi.org/10.1007/978-3-319-46484-8_25

[53] T. Durand, T. Mordan, N. Thome, et M. Cord, (2017). WILDCAT: Weakly Supervised Learning of Deep ConvNets for Image Classification, Pointwise Localization and Segmentation, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, p. 5957-5966. https://doi.org/10.1109/CVPR.2017.631

[54] W. Shimoda et K. Yanai, (2020). Weakly supervised semantic segmentation using distinct class specific saliency maps, *Computer Vision and Image Understanding*, vol. 191, p. 102712. https://doi.org/10.1016/j.cviu.2018.08.006

[55] E. J. Rosana, C. Petitjean, P. Honeine, et F. Abdallah, (2020). BB-UNet: U-Net With Bounding Box Prior, *IEEE J. Sel. Top. Signal Process.*, vol. 14, nº 6, p. 1189-1198. https://doi.org/10.1109/JSTSP.2020.3001502

[56] C. Rother, V. Kolmogorov, et A. Blake, « GrabCut » — Interactive Foreground Extraction using Iterated Graph Cuts, p. 6.

[57] J. Dai, K. He, et J. Sun, (2015). BoxSup: Exploiting Bounding Boxes to Supervise Convolutional Networks for Semantic Segmentation, in *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, p. 1635-1643. https://doi.org/10.1109/ICCV.2015.191

[58] A. Khoreva, R. Benenson, J. Hosang, M. Hein, et B. Schiele, (2017). Simple Does It: Weakly Supervised Instance and Semantic Segmentation, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, p. 1665-1674. https://doi.org/10.1109/CVPR.2017.181

[59] C. Song, Y. Huang, W. Ouyang, et L. Wang, (2019). Box-driven Class-wise Region Masking and Filling Rate Guided Loss for Weakly Supervised Semantic Segmentation, *arXiv:1904.11693 [cs]*. https://doi.org/10.1109/CVPR.2019.00325

[60] J. Lee, J. Yi, C. Shin, et S. Yoon, (2021). BBAM: Bounding Box Attribution Map for Weakly Supervised Semantic and Instance Segmentation, *arXiv:2103.08907 [cs]*. https://doi.org/10.1109/CVPR46437.2021.00267

[61] K.-K. Maninis, S. Caelles, J. Pont-Tuset, et L. Van Gool, (2018). Deep Extreme Cut: From Extreme Points to Object Segmentation, *arXiv:1711.09081 [cs]*. https://doi.org/10.1109/CVPR.2018.00071

[62] D. Lin, J. Dai, J. Jia, K. He, et J. Sun, (2016). ScribbleSup: Scribble-Supervised Convolutional Networks for Semantic Segmentation, *arXiv:1604.05144 [cs]*. https://doi.org/10.1109/CVPR.2016.344

[63] P. Krähenbühl et V. Koltun, (2012). Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials, *arXiv:1210.5644 [cs]*.

[64] P. Vernaza et M. Chandraker, (2017). Learning random-walk label propagation for weakly-supervised semantic segmentation, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 2953-2961. https://doi.org/10.1109/CVPR.2017.315

[65] M. Pu, Y. Huang, Q. Guan, et Q. Zou, (2018). GraphNet: Learning Image Pseudo Annotations for Weakly-Supervised Semantic Segmentation, in *Proceedings of the 26th ACM in-*

ternational conference on Multimedia, Seoul Republic of Korea, p. 483-491. https://doi.org/10.1145/3240508.3240542

[66] A. Bearman, O. Russakovsky, V. Ferrari, et L. Fei-Fei, (2016). What's the Point: Semantic Segmentation with Point Supervision, *arXiv:1506.02106 [cs]*. https://doi.org/10.1007/978-3-319-46478-7_34

[67] R. Qian, Y. Wei, H. Shi, J. Li, J. Liu, et T. Huang, (2019). Weakly Supervised Scene Parsing with Point-Based Distance Metric Learning, *AAAI*, vol. 33, p. 8843-8850. https://doi.org/10.1609/aaai.v33i01.33018843

[68] R. A. McEver et B. S. Manjunath, (2020). PCAMs: Weakly Supervised Semantic Segmentation Using Point Supervision, *arXiv:2007.05615 [cs]*.

[69] J. Ma, (2020). Segmentation Loss Odyssey, *arXiv:2005.13449 [cs, eess]*.

9 Authors

Youssef Ouassit is a PhD student in computer sciences at University Hassan 2, Casablanca, Morocco.

Soufiane Ardchir is a professor of computer sciences at the National School of Marketing and Management, University Hassan 2, Casablanca, Morocco. His area of expertise is big data and machine learning algorithms.

Mohamed Azaouzi is a professor of computer engineering at the University Hassan 2, Casablanca, Morocco. His area of expertise is big data and machine learning algorithms. Dr. Azouazi is a contributing author for more than 50 journal articles and conference papers.

Mohammed Yassine El Ghoumari is a professor of computer sciences at the National School of Marketing and Management, University Hassan 2, Casablanca, Morocco. His area of expertise is big data and machine learning algorithms.

Article submitted 2022-04-08. Resubmitted 2022-05-19. Final acceptance 2022-05-19. Final version published as submitted by the authors.

10 Appendix

Table 8. FCN based weakly supervisied segmentation methods

Source	Model	Mechanism
Pathak et al (2015a)	Constrained CNN	MIL Loss
Saleh et al (2016)		FCN with a foreground/background mask
Lin et al. (2016)	ScribbleSup	Superpixels and graph-cut
Chaudhry et al. (2017)	Fully Convolutional Attention Network (FCAN)	Erasing, attention
Song et al. (2019)	Box-driven classwise masking model	
Cholakkal et al. (2019)	Counting and Segmentation	Classification, density
Huang et al. (2020)		Deep Seeded Region Growing

Table 9. CNN based weakly supervisied segmentation methods (part 1)

Source	Model	Mechanism
Wei et al. (2014)	Hypotheses-CNN-Pooling (HCP)	Shared CNN
Pinheiro and Collobert (2015)		Aggregation layer
Pathak et al. (2015a)		Constraints output loss function
Oquab et al. (2015)		Adaptation layers and multiscale object recognition
Papandreou et al. (2015)	Expectation-Maximization (EM)	Bounding box annotations
Khoreva et al. (2016)		Cross entropy and normalized cut
Kolesnikov and Lampert (2016)	SEC	Seed, expand, constrain and GWRP
Wei et al. (2017)	Adversarial erasing (AE)	AE, online PSL, CAM
Roy and Todorovic (2017)	CRF-RNN	Top-down attention and bottom-up segmentation
Vernaza and Chandraker (2017)	Random-walk Weakly supervised segmentation (RAWKS)	Sparse labels and random-walk hitting probabilities
Kwak et al. (2017)	Superpixel pooling network(SPN)	DeCoupledNet
Hung et al. (2018)	Deep Seeded Region Growing (DSRG)	SRG, CAMs, GAP
Redondo-Cabrera et al. (2018)		Hide and Seek, CAMs, CRF

Table 10. CNN based weakly supervised segmentation methods (part 2)

Source	Model	Mechanism
Tang et al. (2018)		MRF/CRF regularization
Li et al. (2018a)	Guided attention inference network (GAIN)	Attention and Grad-CAM
Wang et al. (2018)	Mining Common Object Features (MCOF)	Seed and bayesian
Li et al. (2018b)		GrabCut, MCG, Grad-CAM, MAP
Chang et al. (Chang et al., 2020)		Superpixels correlation affinities
Wang et al. (Wang et al., 2020)		SEAM method
Chong et al. (Chong et al., 2021)		Two stage training to optimize CAMs
Wan et al. (Wan et al., 2021)	CODNet model and MR-CAM	Co-attention
Xu et al. (2021b)		Atrous Convolutional Feature Network

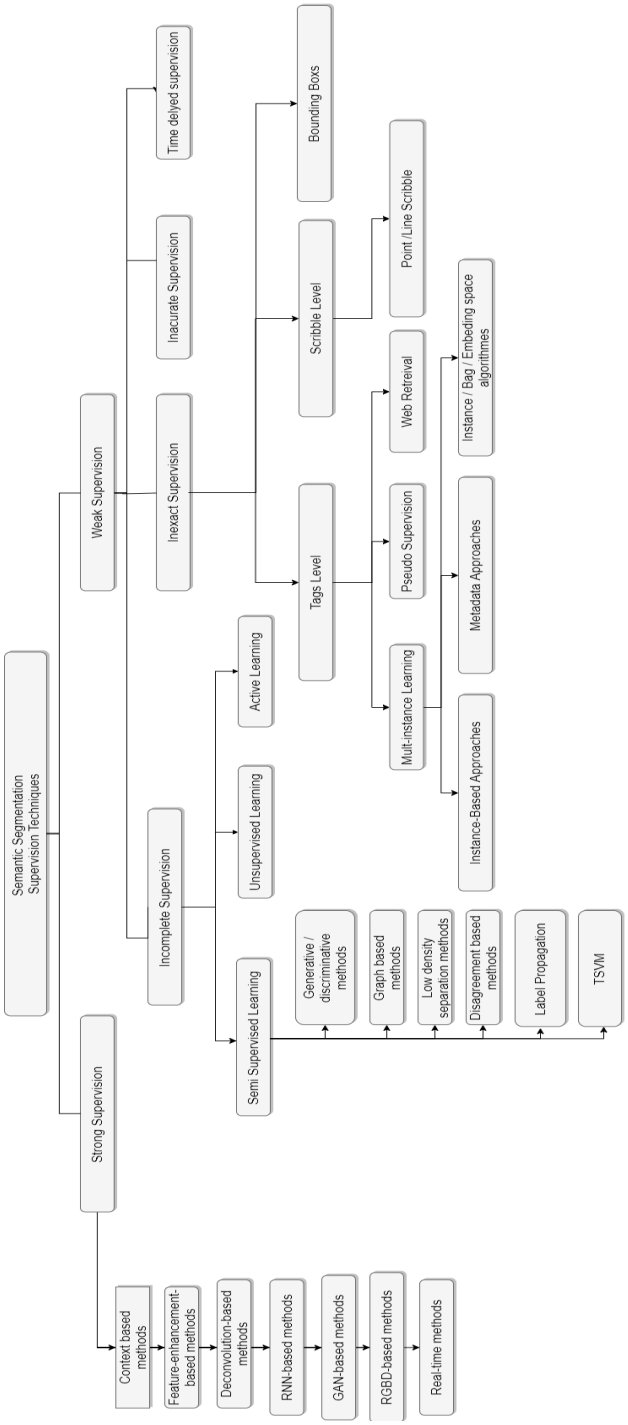

Fig. 15. Taxonomy of semantic segmentation techniques