



A survey of semi- and weakly supervised semantic segmentation of images

Man Zhang^{1,2} · Yong Zhou^{1,2} · Jiaqi Zhao^{1,2} · Yiyun Man³ · Bing Liu^{1,2} · Rui Yao^{1,2}

© Springer Nature B.V. 2019

Abstract

Image semantic segmentation is one of the most important tasks in the field of computer vision, and it has made great progress in many applications. Many fully supervised deep learning models are designed to implement complex semantic segmentation tasks and the experimental results are remarkable. However, the acquisition of pixel-level labels in fully supervised learning is time consuming and laborious, semi-supervised and weakly supervised learning is gradually replacing fully supervised learning, thus achieving good results at a lower cost. Based on the commonly used models such as convolutional neural networks, fully convolutional networks, generative adversarial networks, this paper focuses on the core methods and reviews the semi- and weakly supervised semantic segmentation models in recent years. In the following chapters, existing evaluations and data sets are summarized in details and the experimental results are analyzed according to the data set. The last part of the paper is an objective summary. In addition, it points out the possible direction of research and inspiring suggestions for future work.

Keywords Semi-supervised · Weakly supervised · Semantic segmentation · Review

✉ Yong Zhou
yzhou@cumt.edu.cn

Man Zhang
ts17170006a3@cumt.edu.cn

Jiaqi Zhao
jiaqizhao@cumt.edu.cn

Yiyun Man
man_yy@163.com

Bing Liu
liubing@cumt.edu.cn

Rui Yao
ruiyao@cumt.edu.cn

¹ School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China

² Mine Digitization Engineering Research Center of Ministry of Education of the People's Republic of China, Xuzhou 221116, China

³ Qian Xuesen Laboratory of Space Technology, Beijing 100094, China

1 Introduction

Semantic segmentation is also referred to as a pixel-level classification (Chen and Gupta 2015) because it classifies each pixel of an image into a corresponding class. Semantic segmentation has long been one of the most important tasks in the field of computer vision. It is a commonplace to use deep learning methods to solve semantic segmentation problems. Previously, people used to pay more attention to features and classification methods (Liu et al. 2018). Nowadays, semantic segmentation has a rich application background and has achieved remarkable results. Specifically, segmenting daily life images and natural images (Chen et al. 2019), real-time semantic segmentation of road-driving images (Orsic et al. 2019) and segmentation based on character pose (Zhang et al. 2019). At present, segmenting medical images has received more attention (Goceri and Goceri 2017). Zhao et al. (2019) focused on brain magnetic resonance imaging segmentation. Another study is to introduce block chain miner computation into the field of biomedical image segmentation (Li et al. 2019a). Shen et al. (2017) comprehensively summarized and analyzed the deep learning methods applied to medical image processing. The reason is efficient formulas behind deep learning success (Goceri 2018), significant effect of image based diagnosis systems in biomedical information technology (Goceri and Songul 2018) and future healthcare (Goceri 2017). In addition, the remote sensing image segmentation task has also made a great progress. Mou et al. (2019) enhanced the presentation capabilities of the model used to segment the aerial scene image. Kampffmeyer et al. (2016a) used urban remote sensing images to focus on small objects to map land cover.

From the perspective of supervision, image semantic segmentation methods can be divided into three categories, including fully supervised learning, semi- and weakly supervised learning and unsupervised learning. Unlike pixel-level annotation of fully supervised learning, weakly supervised semantic segmentation tasks hope to use image-level labels in the training process to ultimately predict the object to which each pixel belongs (Vezhn-evets and Buhmann 2010; Yu et al. 2018). The image-level label means that each training image I is represented by a n -dimensional vector set N , and each element of N is regarded as a Boolean variable, and a value of 1 or 0 indicates whether the label exists in the image. To achieve balance, semi-supervised semantic segmentation uses both weak and strong tags in an attempt to compromise the challenge of weak supervision and the high consumption of full supervision (Yu et al. 2018). Unsupervised semantic segmentation, as the name suggests, does not use any labeled data when training deep models (Sultana et al. 2019).

Within the scope of fully supervised segmentation, many excellent models emerged on the basis of CNN. Long et al. (2015) first used the FCN for semantic segmentation. Fast R-CNN was addressed to accelerate Region-based convolutional network (R-CNN) and improve accuracy (Girshick 2015). Later, both faster R-CNN (Ren et al. 2015) and mask R-CNN (He et al. 2018) have been proved to be effective in semantics segmentation. DeepLabv3 (Chen et al. 2017) uses multi-proportional atrous convolution to capture multi-scale information in cascade or in parallel. Atrous convolution, also known as dilated convolution. It is a generalization of the Kronecker factor convolution filter (Zhou et al. 2015), or a traditional convolution using an upsampling filter that supports exponentially extended receive fields without loss of resolution (Garcia-Garcia et al. 2018). DeepLabV3+ (Chen et al. 2018) uses an encoder-decoder architecture to further improve the accuracy and speed of the segmentation algorithm. Despite the gratifying achievements of the fully supervised learning algorithm, this is based on time-consuming and laborious manual annotations. Therefore, semi- and weakly supervised learning algorithm have received more attention. As mentioned in Sect. 2, many semi- and weakly

supervised semantic segmentation models have emerged, and excellent results have been achieved in various fields. More details will be found in Sect. 2. The unsupervised learning algorithm has the lowest level of supervision and does not require any expert annotations. This is a promising direction, but the current results are insufficient and not convincing. The relevant research currently available will be elaborated in the last subsection of the second section.

Many review papers on semantic segmentation have been published. Lateef and Ruichek (2019) gave a thorough review of existing models and data sets. Garcia-Garcia et al. (2018) focused on deep learning techniques with a background of both image and video. What is not refined is that too much space is used to introduce common network architectures and data sets. There are two short reviews (Guo et al. 2017; Siam et al. 2017), and the second one is focused on automated driving. However, Guo et al. (2018) only explained the FCN-based segmentation models, the technical analysis is not sufficient, and there is a lack of review of many weak supervised segmentation methods in the past three years. Similarly, the lack of a review of methods in recent years has also appeared in Yu et al. (2018). Based on classification, Bo et al. (2017) gave a brief introduction about semantic segmentation only in the Sect. 3. Geng et al. (2016) used most of the space to review CNN-based methods. Thoma (2016) gave an analysis of many algorithmic principles and also involves unsupervised learning methods. Moreover, both Garcia-Garcia et al. (2017) and Liu et al. (2018) further analyzed model and data sets. But these papers have a common shortcoming, that is, the lack of integrated analysis of semi- and weakly supervised methods. Zhang et al. (2008) reviewed the unsupervised evaluation method from a novel perspective, please refer to the paper if needed. Vezhnevets and Buhmann (2010) reviewed the weakly supervised semantic segmentation method in the early days, but the paper does not cover the deep learning methods that currently dominate the mainstream. In fact, excellent semi- and weakly supervised learning algorithms emerge in an endless stream, which is also the focus of this paper.

The contributions to this work are summarized as follows:

1. Reviewing the semi- and weakly supervised semantic segmentation models in recent years according to the basic model.
2. Focusing on the algorithm and mechanism of the model and displaying the necessary equations.
3. Comprehensively summarizing the commonly used evaluation indicators and data sets, and then the segmentation effect of different models is analyzed according to the data set.
4. Summarizing the full text and giving inspirational suggestions for future research.

This paper is organized as follows: Sect. 2 reviews semi- and weakly supervised model approaches and concludes with a brief summary of unsupervised methods. Section 3 summarizes the semantic segmentation data sets and various evaluation metrics. Experimental results and analysis are shown in Sect. 4. Section 5 briefly summarizes the paper and lists enlightening suggestions for future study.

2 Models

Whether they are image-level labels, box-level annotations, scribbles or points, it is very cost effective and the main difficulty is how to precisely match these annotations to their corresponding pixels (Huang et al. 2018). This section is a review of semi-supervised and weakly

supervised segmentation methods. According to our knowledge, this is the first review paper on semi- and weakly supervised semantic segmentation in recent years. In the following content, in addition to reviewing the semi-supervised and weakly supervised semantic segmentation model methods, the loss function of the model and the optimization method used are also discussed in depth. Among them, ADAM is nearly a default optimization method for semantic segmentation methods. Chosen scalar products can affect the performance of gradient descent based optimizers (Goceri 2015). Moreover, Sobolev gradient type has been applied recently in some works (Goceri 2016, 2019; Goceri and Esther 2014). At the end of this section, the unsupervised algorithm in recent years will be briefly summarized.

2.1 Semi-supervised methods

Semi-supervised segmentation methods published in recent years are mainly based on three classical models, namely CNN, R-CNN (Girshick et al. 2014) and adversarial networks (Goodfellow et al. 2014).

2.1.1 CNN based models

Based on CNN, Hong et al. (2015) decoupled classification network and segmentation network, and the two parts are trained using image-level and pixel-wise annotations, respectively. In addition, bridging layers are used to output class-specific activation map for obtaining class-specific segmentation maps. Self-supervised method is one of the earliest proposed semi-supervised learning method (Scudder 1965). Zhan et al. (2017) proposed their self-supervised segmentation with a new approach called mix-and-match (M&M) to improve the pre-training. Prior to M&M, the proxy task was proposed to use cross-entropy loss to learn representations from colorization. The tuning loss used to fine-tune the parameters is formulated by transforming from graph optimization to triplet ranking (Schroff et al. 2015), shown as Eq. 1.

$$L = \frac{1}{N} \sum_i \max \left\{ D(P_a^i, P_p^i) - D(P_a^i, P_n^i) + \alpha, 0 \right\} \quad (1)$$

where P_a, P_p, P_n denote anchor, positive, negative nodes in a triplet, α is a regularization factor controlling the distance margin and D is a distance metric measuring patch relationship. Lee et al. (2019) designed the FickleNet by using the core idea of randomly selecting hidden unit, which can also be recognized as a dropout method. In this work, Gradient-weighted Class Activation Mapping (Grad-CAM) proposed by Selvaraju et al. (2016) is used instead of CAM to deal with localization maps, which is expressed as Eq. 2.

$$\text{Grad-CAM}^c = \text{ReLU} \left(\sum_k x_k \times \frac{\partial S^c}{\partial x_k} \right) \quad (2)$$

where x_k is the k th channel of the feature map x , S^c is the classification score of class c .

2.1.2 R-CNN based models

The R-CNN method (Girshick et al. 2014) creatively uses Region Proposal plus CNN framework for target detection and semantic segmentation. The model is shown in Fig. 1. It can be concluded that the workflow of the model is roughly divided into four steps: (a)

getting the input image; (b) extracting candidate regions; (c) entering the candidate regions into the CNN network separately; (d) entering the output of the CNN into the SVM for category determination. Since each region proposal has to perform CNN once, the training process consumes time and space very much. At the same time, some improved models of semi-supervised semantic segmentation based on R-CNN have emerged. Hu et al. (2017) proposed a transfer learning method built on Mask R-CNN (He et al. 2018). Specifically, it is to learn the category-specific mask parameters from the bounding box parameters by a generic weight transfer function shown as Eq. 3.

$$w_{\text{seg}}^c = \mathcal{T}(w_{\text{det}}^c; \theta) \quad (3)$$

where c represents category, w_{det}^c are detection weights in the last layer of the bounding box head, θ is class-agnostic, learned parameters and w_{seg}^c belong to the mask weights.

Mirakhorli and Amindavar (2017) used the hierarchical network structure to interconnect the Mask R-CNN and Conditional Random Field (CRF) to achieve instance segmentation results. The whole structure contains two sub-networks. In sub-network 1, Mask R-CNN is first used to produce object masking and then together with the superpixel layer generated by the original image to produce the final segmentation results. This result enters the Mask R-CNN and superpixel layers in sub-network 2, and the final labeling is generated by the Maximum A Posteriori (MAP) estimated in CRF.

2.1.3 GAN based models

At present, adversarial techniques have received great attention. Although the recently published GAN based semi-supervised semantic segmentation method is not the mainstream method, it has great learning value. The GAN model architecture (Goodfellow et al. 2014) consists of two subnetworks, a generation network and a discriminant network, which can also be called generator (G) and discriminator (D). Details are shown in Fig. 2. During training, G tries to trick D by receiving random noise to generate as realistic a picture as possible, while D is committed to distinguishing between pictures generated by G and real pictures. Thus, D and G constitute a dynamic game process. Instead of using CRF, adversarial method is performed in by Luc et al. (2016) to unify high-order potentials. The designed approach can also be divided into two parts, segmentation network and adversarial network. The RGB image is taken as input in segmentor which produces pixel-wise class predictions. Then adversarial net takes both the output and the RGB image as input to produce final class label. Fischer et al. (2017) further used imperceptible adversarial

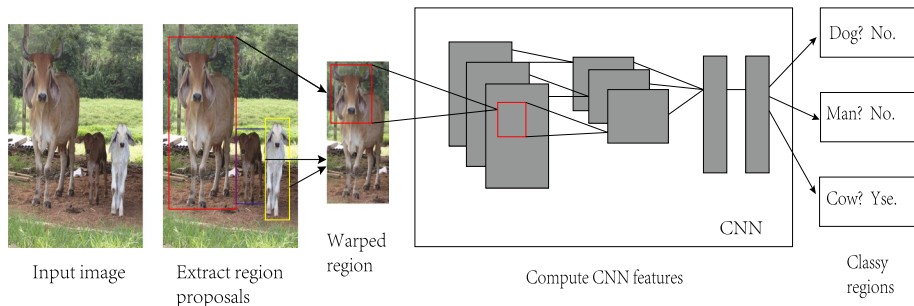


Fig. 1 R-CNN model diagram

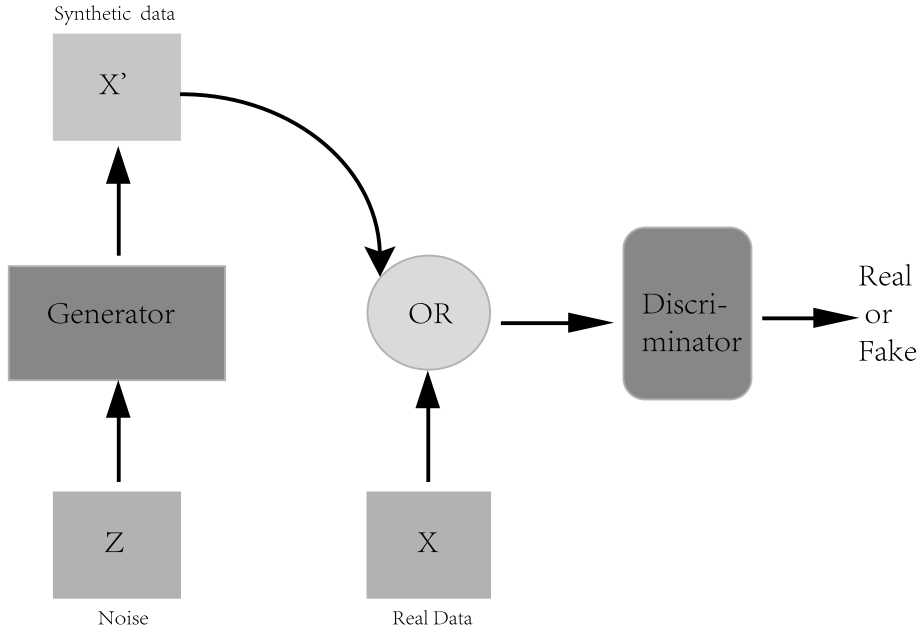


Fig. 2 GAN model flow chart

perturbations to train their model for semantic image segmentation. Different from previous methods where the generators are trained to generate images using noise vectors, the network proposed by Hung et al. (2018) outputs the probability maps of the semantic labels. Under these circumstances, the outputs are enforced close enough to the ground truth label maps spatially. This is fulfilled by combining cross-entropy loss, as shown in Eq. 4.

$$\mathcal{L}_D = - \sum_{h,w} (1 - y_n) \log \left(1 - D(S(\mathbf{X}_n))^{(h,w)} \right) + y_n \log \left(D(\mathbf{Y}_n)^{(h,w)} \right) \quad (4)$$

where $Y_n = 0$ if the sample comes from the segmentation network, $Y_n = 1$ if the sample is from the ground truth label, $D(S(\mathbf{X}_n))^{(h,w)}$ is the confidence map of X at location (h, w) , $D(\mathbf{Y}_n)^{(h,w)}$ is in the same way. As for segmentation network, a multi-class loss is used for training which formed as Eq. 5.

$$\mathcal{L}_{seg} = \mathcal{L}_{ce} + \lambda_{adv} \mathcal{L}_{adv} + \lambda_{semi} \mathcal{L}_{semi} \quad (5)$$

where \mathcal{L}_{ce} , \mathcal{L}_{adv} , \mathcal{L}_{semi} represent spatial multi-class cross entropy loss, adversarial loss, and semi-supervised loss, respectively. All semi-supervised methods mentioned in all three subsections above are summarized in Table 1.

2.2 Weakly supervised methods

It is well known that weakly supervised methods employ different levels of supervision, such as bounding boxes (Dai et al. 2015), scribbles (Lin et al. 2016), points

Table 1 Semi-supervised segmentation methods

Basic model	Source	Model	Main mechanism
CNN	Hong et al. (2015)	Decoupled deep network	Bridge the segmentation and classification
CNN	Zhan et al. (2017)	Mix and match (M&M)	Self-supervision and undirected graph
CNN	Lee et al. (2019)	FickleNet	Randomly selecting hidden unit and Grad-CAM
CNN	Girshick et al. (2014)	R-CNN	Region proposal
R-CNN	Hu et al. (2017)	–	Transfer learning
R-CNN	Mirakhorli and Amindavar (2017)	CNN&CRF	CRF and MAP
GANs	Luc et al. (2016)	Segmentor&Adversarial network	Adversarial training
GANs	Luc et al. (2016)	–	Adversarial perturbations
GANs	Hung et al. (2018)	FCN&GAN	Self-taught and cross-entropy loss

(Russakovsky et al. 2015), image-level labels (Papandreou et al. 2015), etc. The bounding boxes are common representations of object position. Scribbles refer to marking each type of semantics as a mark. Points imply the object location. Among all these types of supervision, image-level supervision is the weakest one and image-level tags can be obtained very efficiently. Although there is a certain gap between models trained by weak supervision and models trained by full supervision, many researchers are devoted to narrowing the gap. Various weakly supervised methods will be elaborated in the rest of this section.

2.2.1 CNN based models

The CNN-based approaches still account for the majority. Only image-level class information is used to train the segmentation model (Pinheiro and Collobert 2014). During training, CNN is used to generate feature planes, and then an aggregation layer takes these planes as input to constrain the model to put more weight on the right pixels. However, one obvious shortcoming of this study is that it only segments the single-object image and unable to meet current needs. Oquab et al. (2015) proposed the idea of transferring the parameters of CNN to overcome other target tasks at an early stage. Papandreou et al. (2015) designed a method called Expectation-Maximization (EM) to train segmentation model under both semi- and weak supervision. The main idea of Hypotheses CNN Pooling (HPC) is each hypothesis fed into CNN produces a c -dimensional prediction and then uses max-pooling for all predictions to get the final multi-target detection (Wei et al. 2014).

Since 2016, a large number of CNN-based weakly supervised semantic segmentation methods have emerged. STC (Wei et al. 2016b) represents a framework for Simple to Complex. It is implemented by three networks: Initial DCNN, Enhanced DCNN and Powerful DCNN, which gradually improve the segmentation performance of the model. Still based on DCNN, Wei et al. (2014) chose Hypotheses-CNN-Pooling (HCP) to predict the classification scores. Additionally, a novel multi-label cross-entropy loss is utilized to work with single-label loss to train the net, shown as Eq. 6.

$$J = -\eta \sum_{i=1}^N \sum_{l=1}^h \sum_{j=1}^w \sum_{m=1}^{c+1} \hat{p}_i^m(i, j) \log(p_i^m(i, j)) \quad (6)$$

where $p_i^m(i, j)$ obtained from the generated localization map is used to represent the groundtruth probability of the m th class at the position (i, j) . Kolesnikov and Lampert (2016) focused on the loss function and proposed a new loss calculation method based on three principles, named SEC. The SEC represents seed, expand, and constrain, respectively. Among them, seeds are localization cues, and seeding loss is used to weakly locate an object. The form of seeding loss is shown as Eq. 7.

$$L_{\text{seed}}(f(X), T, S_c) = -\frac{1}{\sum_{c \in T} |S_c|} \sum_{c \in T} \sum_{u \in S_c} \log f_{u,c}(X) \quad (7)$$

where S_c is a set of locations with label of class c . Given that global max-pooling (GMP) often underestimates object size, and global average-pooling (GAP) often overestimates size. Using global weighted rank-pooling (GWRP) which is leveraged by expansion loss to reasonably extend regions of object seeds. The loss function is as Eq. 8.

$$L_{expand}(f(X), T) = -\frac{1}{|T|} \sum_{c \in T} \log G_c(f(X); d_+) - \frac{1}{|C' \setminus T|} \sum_{c \in C' \setminus T} \log (1 - G_c(f(X); d_-)) - \log G_{c_{bg}}(f(X); d_{bg}) \quad (8)$$

where G_c represents the GWRP classification scores. In addition, a fully connected CRF was designed. Constrain-to-boundary loss is obtained by calculating the mean KL-divergence between the network outputs and the CRF outputs, as shown in Eq. 9.

$$L_{constrain}(X, f(X)) = \frac{1}{n} \sum_{u=1}^n \sum_{c \in C} Q_{u,c}(X, f(X)) \log \frac{Q_{u,c}(X, f(X))}{f_{u,c}(X)} \quad (9)$$

The goal is to make the mask be closer to the boundary. The method proposed in this paper has been used as a reference for subsequent studies (Shen et al. 2018). Similar ideas are applied by Huang et al. (2018) where seeding loss and boundary loss are adapted to obtain better results. Redondo-Cabrera et al. (2018) designed a segmentation model that is fully end-to-end trained and does not require any external aid, such as saliency and priors. The model architecture consists of two parts, the hide-and-seek module and the segmenter module. The hide-and-seek part uses two siamese CAM modules in combination to get activation masks that cover full objects. According to previous activation maps, the segmenter network learns to realize images segmentation using a CNN network. The idea of hide-and-seek is also applied by Singh and Lee (2017). The difference is that this method randomly hides the blocks of images during training, instead of making algorithm changes or relying on external information. Inspired by Weston et al. (2012), Goodfellow et al. (2015), Tang et al. (2018) aimed to evaluate the seeds where labels are known and consistency of all pixels using two evaluation methods. The former uses cross entropy loss and the latter uses normalized cut. It is worth mentioning that normalized Cut is a variant of a family of spectral clustering and embedding algorithms (Shi and Malik 2000; Ng et al. 2001). Moreover, they continue their recent work which entirely uses the normalized cut loss and directly integrates the standard targets in shallow segmentation (Tang et al. 2018). Commonly used proposal generation methods include dense CRF mean-field inference (Papandreou et al. 2015; Rajchl et al. 2016) or graphic cut (Lin et al. 2016). Instead, the proposed method directly combines integrating shallow regularizers with loss functions. There are 2 losses, named Potts loss, CRF loss and kernel cut loss. The joint loss is shown as Eq. 10.

$$\sum_{p \in \Omega_c} H(Y_p, S_p) + \lambda \cdot R(S) \quad (10)$$

where H is the cross entropy between prediction S_p and ground truth labeling Y_p . A segmentation network named guided network (GAIN) was addressed by incorporating the attention (Li et al. 2018a). It contains two routes, S_{cl} and S_{am} . S_{cl} locates the significant area, and S_{am} tries to ensure the coverage accuracy of the area. The final self-guidance loss is as Eq. 11.

$$L_{self} = L_{cl} + \alpha L_{am} \quad (11)$$

Grad-CAM is used here, and it can be concluded that replacing CAM with Grad-CAM has become a new trend. It is proposed that a GrabCut-like algorithm is used to obtain labels from given bounding boxes and achieve the advanced quality through a single training round (Khoreva et al. 2016). The model employs the DeepLabv1 (Chen et al. 2016). Besides, Box-driven figureground segmentation (Rother et al. 2004) and object proposal

(Pont-Tuset and Van Gool 2015) are used to feed the training. MCG (Pont-Tuset et al. 2016) and GrabCut+ are used to mark foreground pixels. A model named CRF-RNN is presented by Roy and Todorovic (2017) where CRF is designed as a Recurrent Neural Network (RNN) and be further used to refine the initial CNN's prediction. This design follows Zhou et al. (2015). The architecture is a totally end to end deep architecture that unifies top-down attention and bottom-up segmentation, and finally refines all the former cues. Similar bottom-up and top-down frameworks are also used by Wang et al. (2018). The proposed MCOF network uses the heat map response of the classification network as the initial seed, and the predictions of RegionNet and PixelNet alternately become the supervision labels of each other and iterate in multiple rounds. For the Refinement step, a Bayesian estimate is used to refine the object area, as shown in Eq. 12.

$$p(obj|v) = \frac{p(obj)p(v|obj)}{p(obj)p(v|obj) + p(bg)p(v|bg)} \quad (12)$$

Set $p(obj)$ as the saliency map, and $p(bg) = 1 - p(obj)$, $p(v|obj)$ and $p(v|bg)$ are feature distributions at object regions and background regions. Vernaza and Chandraker (2017) used sparse labels that are inexpensively available as input to the CNN-based segmentation network, mimicking these tags through training, and ultimately producing dense labels. This method is similar to Lin et al. (2016) but avoids the problem that the upper limit never increases due to non-adaptive label smoothness. The label propagation process is defined by random-walk hitting probabilities (Grady 2006), which is known to be efficiently computed by solving linear systems. Kwak et al. (2017) designed a superpixel pooling network (SPN) and combined with deCoupledNet to perform weakly supervised semantic segmentation tasks. Specifically, SPN is used to generate segmentation annotations, deCoupledNet is used for semantic segmentation. The loss function used to learn SPN is defined by the sum of C binary classification losses, shown as Eq. 13.

$$\mathcal{L}(f(\mathbf{x}), \mathbf{y}) = \frac{1}{C} \sum_{c=1}^C \left\{ y_c \log \frac{e^{f_c(\mathbf{x})}}{1 + e^{f_c(\mathbf{x})}} + (1 - y_c) \log \frac{1}{1 + e^{f_c(\mathbf{x})}} \right\} \quad (13)$$

where $f_c(\mathbf{x})$ and y_c are the network output and the ground-truth label for a single class c . Unlike most existing semantic partitions that focus on countable objects, Li et al. (2018b) not only segments semantics and instances but also splits countable and uncountable categories. In this paper, countable objects and uncountable objects are represented as thing and stuff, respectively, also referred to as panoramic segmentation (Kirillov et al. 2019). The authors assumed that the training data for pixel-level tasks is statistically correlated within an image, and that only small sets of pixels need to be randomly extracted during training. Specifically, the method includes many common mechanisms. For example, GrabCut (Rother et al. 2004) and MCG (Arbeláez et al. 2014) are used to obtain foreground masking, Grad-CAM (Selvaraju et al. 2016) is responsible for positioning tasks, and Maximum-a-Posteriori (MAP) estimate of CRF is the final output.

All of the above CNN-based weak supervised segmentation methods are summarized in Table 2.

2.2.2 FCN based models

Since the FCN was first proposed for semantic segmentation (Long et al. 2015), a large number of weakly supervised methods have been developed. The classic FCN architecture

Table 2 CNN based weakly supervised segmentation methods

Source	Model	Main mechanism
Wei et al. (2014)	Hypotheses-CNN-Pooling (HCP)	Shared CNN
Oquab et al. (2015)	–	Adaptation layers and multiscale object recognition
Papandreou et al. (2015)	Expectation-Maximization (EM)	Boundingbox annotations
Pinheiro and Collobert (2014)	–	Aggregation layer
Kolesnikov and Lampert (2016)	SEC	Seed, expand, constrain and GWRP
Wei et al. (2016a)		
Wei et al. (2016b)		
Hung et al. (2018)	Deep Seeded Region Growing (DSRG)	SRG, CAMs, GAP
Wei et al. (2017)	Adversarial erasing (AE)	AE, online PSL, CAM
Redondo-Cabrera et al. (2018)	–	Hide and Seek, CAMs, CRF
Singh and Lee (2017)		
Tang et al. (2018)	–	MRF/CRF regularization
Li et al. (2018a)	Guided attention inference network (GAIN)	attention and Grad-CAM
Khoreva et al. (2016)		Cross entropy and normalized cut
Roy and Todorovic (2017)	CRF-RNN	Top-down attention and bottom-up segmentation
Wang et al. (2018)	Mining Common ObjectFeatures (MCOF)	Seed and bayesian
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
Li et al. (2018b)	–	GrabCut, MCG, Grad-CAM, MAP

is shown in Fig. 3 and the main difference from CNN is the use of convolution layers instead of fully connected layers. In the same year, several FCN-based weakly supervised semantic segmentation models were released. Russakovsky et al. (2015) used point-level supervision whose supervisory level is slightly higher than image-level supervision. Additionally, a modified training loss function was delivered to solve the difficulty of not being able to learn the full object extent. Details are shown in Eq. 14.

$$\mathcal{L}_{obj}(S, P) = -\frac{1}{|I|} \sum_{i \in I} \left(P_i \log \left(\sum_{c \in \mathcal{O}} S_{ic} \right) + (1 - P_i) \log \left(1 - \sum_{c \in \mathcal{O}} S_{ic} \right) \right) \quad (14)$$

Let P_i be the probability that pixel i belongs to an object. \mathcal{O} be the classes corresponding to objects, with the other classes corresponding to backgrounds. Quab et al. (2015) used a stochastic gradient descent with a global maximum pool, and additionally defined the sum of K binary logistic regression losses as a loss function. Since this is an earlier model, there are problems such as simple structure and weak persuasiveness in the experimental part. Pathak et al. (2015) heuristically defined a multi-class MIL loss, shown as Eq. 15.

$$(x_l, y_l) = \arg \max_{\forall (x, y)} \hat{p}_l(x, y) \quad \forall l \in \mathcal{L}_I \Rightarrow \text{MILLOSS} = \frac{-1}{|\mathcal{L}_I|} \sum_{l \in \mathcal{L}_I} \log \hat{p}_l(x_l, y_l) \quad (15)$$

It is stated that the calculation loss on the largest score pixel is identified only in the rough heat map of the class existing in the image and the background and is propagated back through the network. Let the input image be I , its label set be \mathcal{L}_I , and $\log \hat{p}_l(x_l, y_l)$ be the output heat-map for the l th label at location (x, y) .

After 2015, FCN-based weakly supervised semantic segmentation methods have gained more attention. Adding a separate branch to locate the target object is one of the most commonly used methods (Qi et al. 2016). The proposed localization branch performs as an object detector and help adjust the output of the segmentation branch. The model designed by Lin et al. (2016) has been mentioned many times with good practicability and interactivity. It uses scribbles as annotations and utilizes the methods

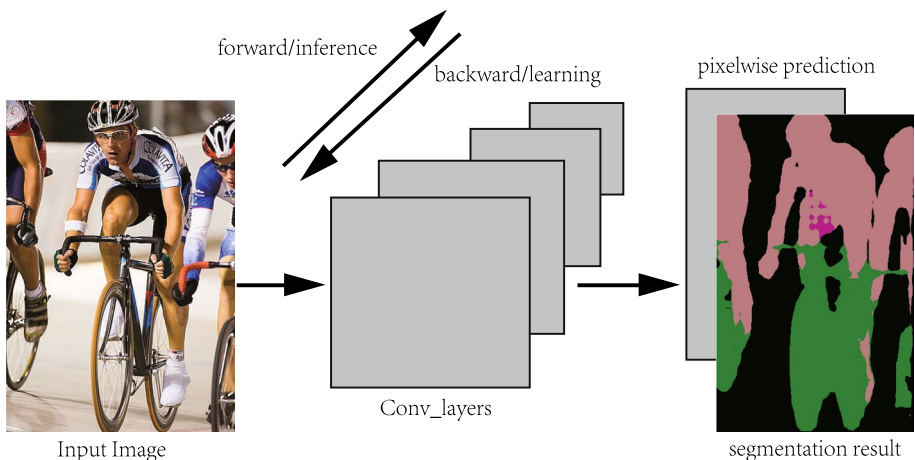


Fig. 3 FCN model diagram

proposed by Felzenszwalb and Huttenlocher (2004) to generate superpixels. It also uses the graph-cut to train the network. The objective function of this method is shown in Eq. 16 which contains two parts, a unary term and a pairwise term.

$$\sum_i \psi_i(y_i|X, S) + \sum_{i,j} \psi_{ij}(y_i, y_j|X) \quad (16)$$

This form of structure has been used in many interactive segmentation methods more than a decade ago (Rother et al. 2004; Grady 2006; Levin et al. 2006; Liu et al. 2009; Jian and Jung 2016). A model called Fully Convolutional Attention Network (FCAN) was designed by Chaudhry et al. (2017) where erasing is used to excavate salient regions hierarchically. In the training process, the attention mechanism for locating the most discriminative region is combined with saliency maps. Compared with another similar method (Wei et al. 2017) which uses erasing to extend the attention map and needs to retrain the attention network after each erasing. However, this method keeps the attention network intact and iteratively erasing to discover new significant areas. While this method compensates for this deficiency and is capable of processing multi-object images. A two-phase learning method is designed by Kim et al. (2017) using SEC as baseline segmentation network, and the network structure consists of two identical sub-networks. The first network is used to locate the most significant region and hide it. The second one is used to continue to find the most significant area which is actually the second significant area. Finally, the target object is segmented. However, this network has two obvious shortcomings, it is not shared or non-end-to-end training. As mentioned earlier, Shen et al. (2018) combined the SEC method with an auxiliary training set for training the segmentor obtained from the net and creates precise pixel-level masks for the training images through the bootstrap process. Specifically, the SEC acts as an initial filter for the target domain and the network domain, respectively, and the web images are used to learn better features. There are many other weakly supervised methods (Hong et al. 2017; Jin et al. 2017; Shen et al. 2018). The idea of dilated convolution (Chen et al. 2014, 2016) is integrated in Wei et al. (2018) to improve the discriminative capability by expanding the receptive field. In this method, CAM is responsible for generating the class-specific localization map for each convolution block and gradually increasing the rate of expansion to search for more target-related areas. The model designed by Zhou et al. (2018) does not combine with popular modules or mechanisms. The main idea is to take the local maximum which is named peaks in a class response map and then calculate them backwards.

At the end of this section, we give a brief review about weakly supervised semantic segmentation methods that perform the counting task simultaneously. Crowd counting is the basis for many complicated tasks such as crowd localization (Shaban et al. 2019), abnormal behavior analysis, and scene monitoring (Gao et al. 2019). Cholakal et al. (2019) used the density map to perform object counting while performing image-level supervised semantic segmentation. The model architecture was designed with two main branches, classification and density. The classification branch is used to determine the presence or absence of an object and to generate a pseudo groundtruth to train the density branch.

Table 3 summarizes all of the FCN-based semantic segmentation methods mentioned in this subsection.

Table 3 FCN based weakly supervised segmentation methods

Source	Model	Mechanism
Russakovsky et al. (2015)	–	Modified training loss
Oquab et al. (2015)	–	Stochastic gradient descent, GMP
Pathak et al. (2015)	Constrained CNN(CCNN)	Multi-class MIL loss
Qi et al. (2016)	–	Augmented feedback and object localization brunch
Lin et al. (2016)	ScribbleSup	Superpixels and graph-cut
Chaudhry et al. (2017)	Fully Convolutional Attention Network (FCAN)	Erasing, attention
Kim et al. (2017)	Two-Phase Learning	SEC, threshold, dense CRF
Shen et al. (2018)	Bidirectional transfer learning	SEC, bootstrap, Grabcut refinement
Wei et al. (2018)	–	Dilated convolution, CAM
Zhou et al. (2018)	Peak Response Maps (PRMs)	Peaks
Cholakkal et al. (2019)	Counting and Segmentation	Classification, density

2.2.3 GAN based models

Although the CNN and FCN based weakly supervised semantic segmentation models occupy half of the country, GAN based methods still have a place. Souly et al. (2017) extended the typical GAN and gradually realized semi-supervised and weakly supervised semantic segmentation. Throughout the architecture, Generator uses both noise and class label to generate an image. The role of discriminator D is to predict the classification confidence of picture pixels which has K classes. Softmax is used to obtain the probability that x belongs to a certain category. It is worth mentioning that in addition to using erasing method on the basis of CNN, Wei et al. (2017) also used an adversarial way (AE) to train neural networks. In addition, inspired by AE, Zhang et al. (2018a) designed adversarial complementary learning (ACoL) method to compensate for the lack of AE, that is, to train several independent classification networks in order to obtain a certain object region. The main part of the architecture is two parallel-classifiers consisting of several convolutions, GAPs and a softmax layer are used to obtain complementary regions of interest. Finally, the results of the two classifiers are fused for outputs.

Table 4 lists the semantic segmentation using adversarial learning methods described in this subsection.

Table 4 GAN based weakly supervised segmentation methods

Source	Model	Mechanism
Souly et al. (2017)	–	Noise
Wei et al. (2017)	Adversarial erasing (AE)	Erasing
Zhang et al. (2018a)	Adversarial complementary learning (ACoL)	Parallel-classifiers, GAPs

2.2.4 Other methods

In addition to the several mainstream methods mentioned above, there are other ways to do with weakly supervised semantic segmentation. Especially in the early days, weakly supervised semantic segmentation has received a lot of attention, but the deep neural network method has not been widely used. Images used to be represented as sets of superpixels (Vezhnevets et al. 2011). With image level labels, the training process is achieved by calculating the distance between the centroids of the two superpixels. In 2012, Vezhnevets et al. (2012) designed a Gaussian process-based algorithm to solve the Bayesian optimization problem, which is how to choose the optimal model. The concrete implementation is realized by using Extremely Randomised Hashing Forest (ERHF), which is capable of mapping almost any feature space into a sparse binary representation. The decision forest is the basis of the Semantic Texton Forest (STF) method (Shotton et al. 2008), and the STF was used as the underlying framework. STF structure was further extended by using geometric context estimation tasks as regularizers. Two years later, a patch alignment-based manifold embedding algorithm and a hierarchical BN was proposed by Zhang et al. (2014), superpixel semantics are finally calculated by voting. Xu et al. (2015) designed a unified framework to handle different types of weak markers, image-level markers, bounding boxes, and scribbles. The method divides training images into n super-pixels and clusters all superpixels using the max-margin clustering (Zhao et al. 2008, 2009). The optimization objective function of the process is shown in Eq. 17.

$$\frac{1}{2}tr(W^T W) + \lambda \sum_{p=1}^n \sum_{c=1}^C \xi(\mathbf{w}_c; \mathbf{x}_p, h_p^c) \quad (17)$$

where W is a feature matrix, each column represents a clustering feature of the category, ξ is the cost of dividing the p th superpixel into class c . Saleh et al. (2016) performed a validation and evaluation of foreground or background masks. Unlike the previous semantic segmentation models which use clean labels, Lu et al. (2017) added noisy annotations. Then, a label noise reduction method emerged as the times require which is realized by a sparse learning model based on $L1$ optimization. For a more detailed theoretical analysis, please refer to the paper. Considering that if there is occlusion between the targets, it is difficult to segment the complete object without additional information. Thus, a saliency model is designed to works in parallel with the segmentation network to provide additional information for image labels (Oh et al. 2017). In the same year, Meng et al. (2017) novelly segmented the components of the target object. The author gave a concept that is different from the object region segmentation, that is, partial level segmentation. The defined energy function clearly shows the structure of the model, shown as Eq. 18.

$$E = E_s + E_c + E_p + E_h \quad (18)$$

Let E_s be a segmentation evaluation for each image to distinguish between foreground and background. E_c is the cosegmentation evaluation, which measures the similarity among foregrounds. E_p is the part consistency evaluation among images. E_h is the assessment of part structure consistency. For more information on cosegmentation, please refer to (Ma and Latecki 2013; Meng et al. 2013).

The last method to be mentioned is essentially a multi-mechanism fusion. There are three steps, image level stage, instance level stage and pixel level stage (Ge et al. 2018). In the whole process, the output of each step is used as the input of the next step, and the first

stages perform the role of multi-evidence fusion, the second step removes the outlier by triplet loss based metric learning and density-based clustering (Rodriguez and Laio 2014) and train a classifier for instance filtering. The last step fuses the former maps and make a final prediction.

2.3 Unsupervised methods

In recent years, unsupervised semantic segmentation methods have received some attention. Sultana et al. (2019) proposed the DCP method, which is capable of background estimation and foreground detection in a variety of challenging real-time environments. In addition, domain adaption is a commonly used method of unsupervised semantic segmentation and the current main method for solving unsupervised domain adaptation is the adversarial learning (Hoffman et al. 2016). Domain adaption is a representative method in migration learning, which aims to improve the performance of the target domain model by using information-rich source domain samples. The source domain has rich supervision information, and the target domain indicates the area where the test sample is located, and there is no label or only a small number of labels. Murez et al. (2017) aimed to design an unsupervised domain adaptation framework that is widely applicable and in the field of image processing. In addition, the training process is performed by adding additional networks and losses, as shown in Eq. 19.

$$Q = \lambda_c Q_c + \lambda_z Q_z + \lambda_{tr} Q_{tr} + \lambda_{id} Q_{id} + \lambda_{cyc} Q_{cyc} + \lambda_{trc} Q_{trc} \quad (19)$$

For specific individual loss functions, please refer to the paper as needed. A dual channel-wise alignment networks (DCAN) model was designed by Wu et al. (2018). The author assumed that channel alignment is important for adjusting the segmentation model because it preserves the spatial structure and semantic concepts, thus effectively constraining the domain shift. Saito et al. (2018) introduced a new kind of confrontational learning. The specific implementation is to design two classifiers, which are used to maximize the difference of the target samples to detect the target domain samples far from the source domain, and then generate features that minimize the difference to generate the target domain features close to the source domain. Thereby optimizing the boundary segmentation and aligning the distribution of the source and target domains. Fully Convolutional Adaptation Networks (FCAN) was presented by Zhang et al. (2018b) combined with Appearance Adaptation Networks (AAN) and Representation Adaptation Networks (RAN). The purpose of the ANN network is to obtain high-level content in the source image and low-level pixel information of the target domain. The FCN network is shared in the RAN, and atrous spatial pyramid pooling (ASPP) is additionally used to expand the receptive field of the filter in the feature map. Li et al. (2019b) designed a bidirectional learning system that alternately learns the segmentation adaptive model and the image translation model. The self-supervised learning (SSL) algorithm is used to train the segmentation adaptation model with a new perceptual loss. Then, through the reverse learning, a better segmentation adaptation model will help to obtain a better translation model.

3 Evaluation metrics and datasets

This section describes the existing evaluation metrics and data sets, paving the way for the experiment analysis of the next chapter.

3.1 Evaluation metrics

In order to fairly measure the contribution of the segmentation model approach, the assessment requires the use of standard, accepted methods. Execution time, memory usage, and accuracy are all evaluations. However, due to the different design goals of the model, some indicators will be more convincing than other indicators, so it is necessary to analyze the specific situation. Commonly used evaluation metrics are intersection over union IoU , mean intersection-over-union $mIoU$, average precision AP_{vol}^r , mean average precision mAP , panoptic quality PQ , average best overlap ABO and mean accuracy $mAcc$. Their descriptions are listed as follows.

$$IoU = \frac{\text{Area of overlap}}{\text{Area of union}} \quad (20)$$

$$mIoU = \frac{1}{k+1} \sum_{i=0}^k \frac{P_{ii}}{\sum_{j=0}^k P_{ij} + \sum_{j=0}^k P_{ji} - P_{ii}} \quad (21)$$

$$AP_{vol}^r = \int_0^1 p(r) dr \quad (22)$$

$$mAP = \int_0^1 P(R) dR \quad (23)$$

$$PQ = \frac{\sum_{(p,g) \in TP} IoU(p,g)}{|TP|} \times \frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|} \quad (24)$$

$$ABO = \frac{1}{|G^c|} \sum_{g_i^c \in G^c} \max_{l_j \in L} \text{Overlap}(g_i^c, l_j) \quad (25)$$

$$(g_i^c, l_j) = \frac{\text{area}(g_i^c) \cap \text{area}(l_j)}{\text{area}(g_i^c) \cup \text{area}(l_j)} \quad (26)$$

$$MPA = \frac{1}{k+1} \sum_{i=0}^k \frac{P_{ii}}{\sum_{j=0}^k P_{ij}} \quad (27)$$

3.2 Datasets

CityScapes dataset The Cityscapes dataset (Cordts et al. 2016), the Urban Landscape Dataset, is a large-scale dataset. It contains a set of stereo video sequences that record street scenes in 50 different cities. In addition to a large set of 20,000 weakly annotated frames,

it also contains 5000 frames of high quality pixel-level annotations. The Cityscapes dataset has two evaluation criteria, fine and coarse. The former corresponds to 5000 finely labeled images, while the latter corresponds to 5000 fine labels plus 20,000 rough labels.

Microsoft COCO dataset Microsoft COCO (Lin et al. 2014) is a data set collected by the Microsoft team for image processing. There are five types of tags: target detection, key point detection, object segmentation, polygon segmentation and image description. These tag data are stored in json format. In addition, the COCO dataset has more than 300,000 images, more than 2 million instances, more than 70 categories and multiple objects in each image.

Pascal VOC dataset The PASCAL VOC Challenge mainly includes subclasses such as object classification, object detection, object segmentation, human layout, and action classification. The data set includes JPEG images, annotations, imagesets, Segmentationobject and segmentationclass. JPEG images contain all the images provided by PASCAL VOC, including training images and test images. Annotations mainly stores label files in xml format, and each xml corresponds to a picture in JPEG image. Imagesets includes action, layout, main, and segmentation, where segmentation stores the data for segmentation. Segmentationobject and segmentationclass are used to save the segmentation data. PASCAL VOC 2007 contains 9,963 labeled images with a total of 24,640 objects. The trainval/test of PASCAL VOC 2012 (Everingham et al. 2012) contains all the corresponding pictures from PASCAL VOC 2008 to PASCAL VOC 2010. In trainval, there are 11,540 images with a total of 27,450 objects. For the segmentation task, the trainval of VOC2012 contains all the corresponding pictures from PASCAL VOC 2007 to PASCAL VOC 2011, and test only contains the corresponding pictures from PASCAL VOC 2008 to PASCAL VOC 2011.

4 Experimental comparison and analysis

This chapter summarizes and analyzes the semi- and weakly supervised semantic segmentation algorithms in recent years according to the data set. The following summarizes the three data sets that are used more frequently, namely CityScapes dataset, Microsoft COCO dataset, and Pascal VOC 2012 dataset.

4.1 Pascal Voc 2012 dataset

As we all know, Pascal VOC 2012 is the most commonly used dataset in image processing and even semantic segmentation. The experimental results of methods using this dataset are shown in Table 5. Due to the numerous methods of experimenting with VOC data sets, only the results based on the two most commonly used networks VGG16 and Resnet are shown here. Similarly, the two most commonly used and currently most representative evaluation metrics val-mIoU and test-mIoU are used. From the experimental results of Lee et al. (2019) and Chaudhry et al. (2017) we can conclude that the same mechanism can achieve different effects on different DeepLabs. For example, using ResNet will be better than using VGGNet. The effect of the Dropout Rate and the degree of supervision on the experimental results is additionally given in the paper, and the description will not be repeated here. The implementation of Shen et al. (2018) is based on MXNet (Chen et al. 2015). Tang et al. (2018) made CRF loss a universal performance improvement mechanism that can work effectively on several networks.

Table 5 Results on Pascal Voc 2012 dataset

Source	Backbone	Mode	Val-mIoU	Test-mIoU
Souly et al. (2017)	VGG16	Baseline	59.5	–
		Semi	64.1	–
		Weak	65.8	–
Zhan et al. (2017)	VGG16	Random	56.7	–
		Colorize	64.5	–
Lee et al. (2019)	VGG16	–	61.2	61.9
Shen et al. (2018)	VGG16	–	58.8	60.2
Chaudhry et al. (2017)	VGG16	noCRF	56.5	57.04
		CRF	58.6	59.24
Li et al. (2018b)	VGG16	–	55.3	56.8
Tang et al. (2018)	VGG16	–	64.4	–
Wang et al. (2018)	VGG16	–	56.2	57.6
Lee et al. (2019)	Resnet	–	64.9	65.3
Shen et al. (2018)	Resnet	–	63.0	63.9
Chaudhry et al. (2017)	Resnet	noCRF	59.3	60.3
Tang et al. (2018)	Resnet	–	72.9	–
Zhou et al. (2018)	Resnet	–	53.4	–
Wang et al. (2018)	Resnet	–	60.3	61.2

4.2 CityScapes dataset

This section compares the methods using the CityScapes dataset. The experimental comparison results are shown in Table 6. It can be seen from the results of Hung et al. (2018) that increasing the loss term can improve the experimental results, and this conclusion is also in line with the experimental comparison results of the previous section. As can be drawn from Zhan et al. (2017), after fine-tune, the performance of the model is significantly improved. However, the test result of Li et al. (2019a) is not satisfactory.

4.3 Microsoft CoCo dataset

Finally, the experimental results on the MS COCO dataset are analyzed, as shown in Tables 7, 8 and 9. Comparing the results, it can be found that combining the COCO dataset with other datasets can achieve better training results than simply using the COCO dataset. And this result is almost better than the experimental results of all the above-mentioned datasets that are used alone. In addition to comparing the weak supervised results, these results are compared with the fully supervised ones. It can be discovered that the gap between the weakly supervised and the fully supervised is very small.

From the above comparison, we can get three conclusions very intuitively. First of all, the semi- and weakly supervised learning semantic segmentation field has produced a lot of methods and achieved satisfactory results. Second, the results of weak supervised learning have been comparable to those obtained by full-supervised learning under the same method. Third, focusing on mIoU, it can be found that most of the values are between 50 and 60%. Although the current results are satisfactory, they still have a great distance from

Table 6 Results on CityScapes validation set

Source	Backbone	mIoU	IoU-weak	IoU-full	AP^r_{vol}	AP^r_{vol} -Test	PQ
Hung et al. (2018)-baseline	Adversarial network	66.4	–	–	–	–	–
Hung et al. (2018)-baseline+Ladv	Adversarial network	67.7	–	–	–	–	–
Li et al. (2019a)-thing	PSPNet/ImageNet	–	68.2	70.4	17.0	35.8	12.8
Li et al. (2019a)-stuff	PSPNet/ImageNet	–	60.2	72.4	33.1	43.9	–
Li et al. (2019a)-all	PSPNet/ImageNet	–	63.6	71.6	26.3	40.5	–
Zhan et al. (2017)-baseline	VGG	Random-pretrain:42.5 Colorize-pretrain:57.5	–	–	–	–	–
Zhan et al. (2017)-M&M	VGG-16	49.1	66.4	–	–	–	–

Table 7 Results on Voc2012+COCO

Source	Backbone	Method	mIoU	ABO	mAP _{0.5}	mAP _{0.75}	Test-mIoU
Kwak et al. (2017)	VGG-16	Semi-MnG+	68.9	–	–	–	69.9
	VGG-16	weak-MnG+	71.6	–	–	–	72.8
	VGG-16	full	71.6	–	–	–	73.2
	DeepLabv2+ ResNet101	weak-MnG+	74.2	–	–	–	–
	DeepLabv2+ ResNet101	full	77.7	–	–	–	–
	DeepLabv2 VGG-16	Weak-DeepMask	–	48.8	42.9	11.5	–
	DeepLabv2 VGG-16	Weak-DeepLab-BOX	–	51.41	46.4	18.5	–

Table 8 Results on Pascal VOC and COCO

Source	VOC	COCO	IoU	AP _{vol}	PQ
Li et al. (2019a)	Weak	Weak	75.7	55.5	59.5
	Weak	Full	75.8	56.1	59.8
	Full	Weak	77.5	58.9	62.7
	Full	Full	79.0	59.5	63.1

Table 9 Results on MS COCO 2014 validation

Source	Method	MAP
Zhou et al. (2018)	W/o peak stimulation	53.1
	Full approach	57.5

the ideal ones, and the method of significantly improving the weakly supervised segmentation remains to be explored.

4.4 Segmentation results

This section visualizes the image segmentation effects of some classic semi and weakly supervised methods. Then give some brief analysis according to the segmentation effect of different methods. Table 10 shows the phased segmentation effect of point-level segmentation (Russakovsky et al. 2015). From the segmentation results, it can be concluded that the use of point-level supervision can successfully segment objects in the picture. However, the segmentation effect is rough and the edges are not detailed enough. The performance of decoupled deep neural network with different examples is shown in Table 11. It can be clearly seen that although the effect is not fine enough, the segmentation of the object edges is more accurate. Besides, it is capable of recognizing and segmenting small objects. Table 12 shows the segmentation effect of the semi-supervised semantic segmentation using adversarial learning network (Huang et al. 2018) under different loss functions. The results reflect a significant increase in the performance of this semi-supervised model using the adversarial method compared to earlier years. However, it can be found that although

Table 10 Point-level segmentation (Russakovsky et al. 2015)



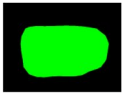
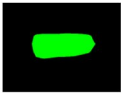
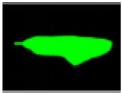
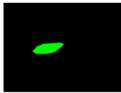







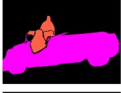




























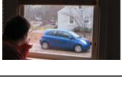





Original image	Ground truth	Img	Img+Obj	Img+Obj+1Point	Full
					
					
					
					

Table 11 Decoupled deep neural network (Hong et al. 2015)








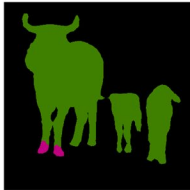


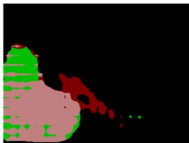
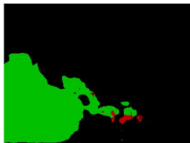




Input image	Ground-truth	5 examples	10 examples	25 examples	Full annotations
					
					
					
					

the segmentation effect is satisfactory on a simple graph, the segmentation result is still poor for complex ones, especially when the objects in the graph are overlapped and interlaced. In addition, batch size is an important factor affecting performance of all deep learning based image segmentation (Goceri and Gooya 2018). Therefore, it should be chosen carefully.

5 Inspiration and conclusion

This paper reviews the semi-supervised and weakly supervised segmentation model methods, focusing on the core content of the model architecture, working mechanism and main functions. In general, although there has been a long-term research on semi-supervised

Table 12 Adversarial learning model (Hung et al. 2018)

Image	Annotation	+Ladv	+Ladv+Lsemi
			
			
			
			

and weakly supervised learning applications in image segmentation, the number of studies in this area 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. It can be seen from this paper that the research on semi-supervised and weakly supervised segmentation methods has made great progress. Many studies have pushed single object semantic segmentation to multi-objective instance segmentation, and even panoramic segmentation and counting. 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.

1. Although some methods such as adding additional mechanisms or designing the loss function can improve the performance of the segmentation model, the results obtained by current methods are still far from the ideal state. Therefore, the next study should focus on two aspects, one is to continuously reduce the degree of supervision, and the other is to continuously improve the segmentation effect while achieving more complex tasks.
2. It can be seen from the use of data sets that the current semi-supervised and weakly supervised semantic segmentation often takes the natural life scene as the application

background, that is to say, there is a problem that the application background is relatively simple. Drawing on the rich application background of fully supervised semantic segmentation, such as medical images (Zhao et al. 2009; Li et al. 2019a), remote sensing images (Zhou et al. 2019; Liu et al. 2019), and so on. Subsequent research focuses on how to utilize weakly supervised semantic segmentation for a wider variety of tasks. Although some studies have now used weakly supervised methods to segment medical images (Jia et al. 2017; Rajchl et al. 2016), it still takes a lot of effort to accurately segment the complex medical images with a small number of annotations.

3. Semantic segmentation is one of the basic tasks of remote sensing image processing. In other words, semantic segmentation of remote sensing images has great research value and practical application significance. Many studies have used deep neural networks to segment remote sensing images (Kampffmeyer et al. 2016b; Wang et al. 2017; Zhang et al. 2017; Hamaguchi et al. 2017). However, fully supervised learning is used in current remote sensing image segmentation methods. Therefore, the use of weakly supervised learning instead of fully supervised learning can effectively solve problems such as the current remote sensing image datasets are not abundant, and the resource waste of collecting pixel by pixel annotations.
4. As mentioned at the end of Sect. 2.2.2, counting can be done simultaneously with segmentation. Therefore, we reasoned that we can implement other related tasks while performing semantic segmentation, such as target behavior recognition and text interpretation, replacing some specified segmentation objects with other objects to generate new images, and so on. In general, subsequent research can consider giving it more practical value on the basis of segmentation.
5. 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 segmentation 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.

Funding Funding was provided by State's Key Project of Research and Development Plan of China (Grant No. 2016YFC0600900), National Natural Science Foundation of China (Grant No. 61572505, 61772530, 61806206), Six Talent Peaks Project in Jiangsu Province (Grant Nos. 2015-DZXX-010, 2018-XYDXX-044), Natural Science Foundation of Jiangsu Province (Grant Nos. BK20180639, BK20171192, BK20180174), China Postdoctoral Science Foundation (Grant No. 2018M642359) and Innovative Research Group Project of the National Natural Science Foundation of China (Grant No. 61801198).

References

- Arbeláez PA, Pont-Tuset J, Barron JT, Marqués F, Malik J (2014) Multiscale combinatorial grouping. In: 2014 IEEE conference on computer vision and pattern recognition, pp 328–335
- Bo Z, Feng J, Xiao W, Yan S (2017) A survey on deep learning-based fine-grained object classification and semantic segmentation. *Int J Autom Comput* 14(2):119–135
- Chaudhry A, Dokania PK, Torr PHS (2017) Discovering class-specific pixels for weakly-supervised semantic segmentation. [arXiv:1707.05821](https://arxiv.org/abs/1707.05821)
- Chen X, Gupta A (2015) Weakly supervised learning of convolutional networks. In: 2015 IEEE international conference on computer vision (ICCV), pp 1431–1439

- Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2014) Semantic image segmentation with deep convolutional nets and fully connected crfs. *Comput Sci* 4:357–361
- Chen T, Li M, Li Y, Lin M, Wang N, Wang M, Xiao T, Xu B, Zhang C, Zhang Z (2015) Mxnet: a flexible and efficient machine learning library for heterogeneous distributed systems. [arXiv:1512.01274](https://arxiv.org/abs/1512.01274)
- Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2016) Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans Pattern Anal Mach Intell* 40:834–848
- Chen LC, Papandreou G, Schroff F, Adam H (2017) Rethinking atrous convolution for semantic image segmentation. [arXiv abs/1706.05587](https://arxiv.org/abs/1706.05587)
- Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder–decoder with atrous separable convolution for semantic image segmentation. In: *Proceedings of the European conference on computer vision (ECCV)*, pp 801–818
- Chen K, Pang J, Wang J, Xiong Y, Li X, Sun S, Feng W, Liu Z, Shi J, Ouyang W, et al. (2019) Hybrid task cascade for instance segmentation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 4974–4983
- Cholakkal H, Sun G, Khan FS, Shao L (2019) Object counting and instance segmentation with image-level supervision. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 12397–12405
- 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) Boxesup: exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: *2015 IEEE international conference on computer vision (ICCV)*, pp 1635–1643
- Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A (2012) The Pascal visual object classes challenge 2012 (VOC2012) results. <http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html>. Accessed 4 Dec 2019
- Felzenszwalb PF, Huttenlocher DP (2004) Efficient graph-based image segmentation. *Int J Comput Vis* 59:167–181
- Fischer V, Kumar MC, Metzen JH, Brox T (2017) Adversarial examples for semantic image segmentation. [arXiv:1703.01101](https://arxiv.org/abs/1703.01101)
- Gao C, Yao R, Zhao J, Zhou Y, Hu F, Li L (2019) Structure-aware person search with self-attention and online instance aggregation matching. *Neurocomputing* 369:29–38
- Garcia-Garcia A, Orts S, Oprea S, Villena-Martinez V, Rodríguez JG (2017) A review on deep learning techniques applied to semantic segmentation. [arXiv:1704.06857](https://arxiv.org/abs/1704.06857)
- Garcia-Garcia A, Orts-Escolano S, Oprea S, Villena-Martinez V, Martinez-Gonzalez P, Garcia-Rodriguez J (2018) A survey on deep learning techniques for image and video semantic segmentation. *Appl Soft Comput* 70:41–65
- Ge W, Yang S, Yu Y (2018) Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning. In: *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp 1277–1286
- Geng Q, Zhou Z, Cao X (2016) Survey of recent progress in semantic image segmentation with cnns. *Sci China Inf Sci* 61:1–18
- Girshick RB (2015) Fast r-cnn. In: *2015 IEEE international conference on computer vision (ICCV)*, pp 1440–1448
- Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 580–587
- Goceri E (2015) Effects of chosen scalar products on gradient descent algorithms. In: *The 28th international conference of the Jangjeon mathematical society (ICJMS)*, Antalya, Turkey, p 115
- Goceri E (2016) Fully automated liver segmentation using sobolev gradient-based level set evolution. *Int J Numer Methods Biomed Eng* 32:e02765
- Goceri E (2017) Future healthcare: will digital data lead to better care? In: *4th world conference on health sciences (HSCI-2017)*, Antalya, Turkey
- Goceri E (2018) Formulas behind deep learning success. In: *International conference on applied analysis and mathematical modeling (ICAAMM)*, Istanbul, Turkey, p 156
- Goceri E (2019) Diagnosis of alzheimer’s disease with sobolev gradient based optimization and 3d convolutional neural network. *Int J Numer Methods Biomed Eng* 35(7):e3225

- Goceri E, Esther DM (2014) A level set method with sobolev gradient and haralick edge detection. In: The 4th world conference on information technology (WCIT 2013), November 26–27, 2013, Brussels, Belgium, vol 5, pp 131–140
- Goceri E, Goceri N (2017) Deep learning in medical image analysis: recent advances and future trends. In: 11th international conferences on computer graphics, visualization, computer vision and image processing (CGVCVIP), pp 305–311
- Goceri E, Gooya A (2018) On the importance of batch size for deep learning. In: International conference on mathematics (ICOMATH), an Istanbul meeting for world mathematicians, minisymposium on approximation theory & minisymposium on math education, Istanbul, Turkey
- Goceri E, Songul C (2018) Biomedical information technology: image based computer aided diagnosis systems. In: International conference on advanced technologies, Antalya, Turkey, p 132
- Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville AC, Bengio Y (2014) Generative adversarial nets. In: NIPS, pp 2672–2680
- Goodfellow IJ, Bengio Y, Courville AC (2015) Deep learning. *Nature* 521:436–444
- Grady L (2006) Random walks for image segmentation. *IEEE Trans Pattern Anal Mach Intell* 28:1768–1783
- Guo Y, Liu Y, Georgiou T, Lew MS (2017) A review of semantic segmentation using deep neural networks. *Int J Multimed Inf Retr* 7:87–93
- Guo Y, Liu Y, Georgiou T, Lew MS (2018) A review of semantic segmentation using deep neural networks. *Int J Multimed Inf Retr* 7(2):87–93
- Hamaguchi R, Fujita A, Nemoto K, Imaizumi T, Hikosaka S (2017) Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery. In: 2018 IEEE winter conference on applications of computer vision (WACV), pp 1442–1450
- He K, Gkioxari G, Dollár P, Girshick RB (2018) Mask r-cnn. *IEEE Int Conf Comput Vision* 2017:2961–2969
- Hoffman J, Wang D, Yu F, Darrell T (2016) Fcns in the wild: pixel-level adversarial and constraint-based adaptation. [arXiv:1612.02649](https://arxiv.org/abs/1612.02649)
- Hong S, Noh H, Han B (2015) Decoupled deep neural network for semi-supervised semantic segmentation. In: Advances in neural information processing systems, pp 1495–1503
- Hong S, Yeo D, Kwak S, Lee H, Han B (2017) Weakly supervised semantic segmentation using web-crawled videos. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 2224–2232
- Hu R, Dollár P, He K, Darrell T, Girshick RB (2017) Learning to segment every thing. In: 2018 IEEE/CVF conference on computer vision and pattern recognition, pp 4233–4241
- Huang Z, Wang X, Wang J, Liu W, Wang J (2018) Weakly-supervised semantic segmentation network with deep seeded region growing. In: 2018 IEEE/CVF conference on computer vision and pattern recognition, pp 7014–7023
- Hung WC, Tsai YH, Liou YT, Lin YY, Yang MH (2018) Adversarial learning for semi-supervised semantic segmentation. In: BMVC
- Jia Z, Huang X, Chang EIC, Xu Y (2017) Constrained deep weak supervision for histopathology image segmentation. *IEEE Trans Med Imaging* 36:2376–2388
- Jian M, Jung C (2016) Interactive image segmentation using adaptive constraint propagation. *IEEE Trans Image Process* 25(3):1301–1311
- Jin B, Segovia MVO, Stsstrunk S (2017) Webly supervised semantic segmentation. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 1705–1714
- Kampffmeyer M, Salberg AB, Jenssen R (2016a) Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 1–9
- Kampffmeyer M, Salberg AB, Jenssen R (2016b) Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In: 2016 IEEE conference on computer vision and pattern recognition workshops (CVPRW), pp 680–688
- Khoreva A, Benenson R, Hosang JH, Hein M, Schiele B (2016) Simple does it: weakly supervised instance and semantic segmentation. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 1665–1674
- Kim D, Cho D, Yoo D, Kweon IS (2017) Two-phase learning for weakly supervised object localization. In: 2017 IEEE international conference on computer vision (ICCV), pp 3554–3563
- Kirillov A, He K, Girshick R, Rother C, Dollár P (2019) Panoptic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 9404–9413
- Kolesnikov A, Lampert CH (2016) Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In: European conference on computer vision. Springer, pp 695–711

- Kwak S, Hong S, Han B (2017) Weakly supervised semantic segmentation using superpixel pooling network. In: AAAI
- Lateef F, Ruichek Y (2019) Survey on semantic segmentation using deep learning techniques. *Neuro-computing* 338:321–348
- 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
- Levin A, Lischinski D, Weiss Y (2006) A closed-form solution to natural image matting. *IEEE Trans Pattern Anal Mach Intell* 30:228–242
- Li K, Wu Z, Peng KC, Ernst J, Fu Y (2018a) Tell me where to look: guided attention inference network. In: *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp 9215–9223
- Li Q, Arnab A, Torr PHS (2018b) Weakly- and semi-supervised panoptic segmentation. In: *ECCV*, pp 102–118
- Li B, Chenli C, Xu X, Jung T, Shi Y (2019a) Exploiting computation power of blockchain for biomedical image segmentation. In: *The IEEE conference on computer vision and pattern recognition (CVPR) workshops*
- Li Y, Yuan L, Vasconcelos N (2019b) Bidirectional learning for domain adaptation of semantic segmentation. In: *The IEEE conference on computer vision and pattern recognition (CVPR)*, pp 6936–6945
- 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 D, Dai J, Jia J, He K, Sun J (2016) Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In: *2016 IEEE conference on computer vision and pattern recognition (CVPR)*, pp 3159–3167
- Liu J, Sun J, Yeung Shum H (2009) Paint selection. In: *SIGGRAPH 2009*:69
- Liu X, Deng Z, Yang Y (2018) Recent progress in semantic image segmentation. *Artif Intell Rev* 52(2):1089–1106
- Liu X, Zhou Y, Zhao J, Yao R, Liu B, Zheng Y (2019) Siamese convolutional neural networks for remote sensing scene classification. *IEEE Geosci Remote Sens Lett* 16:1–5
- 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
- Lu Z, Xiang T, Han P, Wang L, Gao X (2017) Learning from weak and noisy labels for semantic segmentation. *IEEE Trans Pattern Anal Mach Intell* 39(3):486–500
- Luc P, Couprie C, Chintala S, Verbeek J (2016) Semantic segmentation using adversarial networks. [arXiv:1611.08408](https://arxiv.org/abs/1611.08408)
- Ma T, Latecki LJ (2013) Graph transduction learning with connectivity constraints with application to multiple foreground cosegmentation. In: *2013 IEEE conference on computer vision and pattern recognition*, pp 1955–1962
- Meng F, Li H, Ngan KN, Zeng L, Wu Q (2013) Feature adaptive co-segmentation by complexity awareness. *IEEE Trans Image Process* 22:4809–4824
- Meng F, Li H, Wu Q, Luo B, Ngan KN (2017) Weakly supervised part proposal segmentation from multiple images. *IEEE Trans Image Process* 26:4019–4031
- Mirakhorli J, Amindavar H (2017) Semi-supervised hierarchical semantic object parsing. In: *2017 3rd Iranian conference on intelligent systems and signal processing (ICSPIS)*, pp 48–53
- Mou L, Hua Y, Zhu XX (2019) A relation-augmented fully convolutional network for semantic segmentation in aerial scenes. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 12416–12425
- Murez Z, Kolouri S, Kriegman DJ, Ramamoorthi R, Kim K (2017) Image to image translation for domain adaptation. In: *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp 4500–4509
- Ng AY, Jordan MI, Weiss Y (2001) On spectral clustering: analysis and an algorithm. In: *NIPS*, pp 849–856
- Oh SJ, Benenson R, Khoreva A, Akata Z, Fritz M, Schiele B (2017) Exploiting saliency for object segmentation from image level labels. In: *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp 5038–5047
- Oquab M, Bottou L, Laptev I, Sivic J (2015) Is object localization for free? Weakly-supervised learning with convolutional neural networks. In: *2015 IEEE conference on computer vision and pattern recognition (CVPR)*, pp 685–694
- Orsic M, Kreso I, Bevandic P, Segvic S (2019) In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 12607–12616

- Papandreou G, Chen LC, Murphy K, Yuille AL (2015) Weakly- and semi-supervised learning of a dcnn for semantic image segmentation. In: Proceedings of the IEEE international conference on computer vision 1742–1750
- Pathak D, Krähenbühl P, Darrell T (2015) Constrained convolutional neural networks for weakly supervised segmentation. In: 2015 IEEE international conference on computer vision (ICCV), pp 1796–1804
- Pinheiro PHO, Collobert R (2014) From image-level to pixel-level labeling with convolutional networks. In: 2015 IEEE conference on computer vision and pattern recognition (CVPR), pp 1713–1721
- Pont-Tuset J, Van Gool L (2015) Boosting object proposals: from pascal to coco. In: Proceedings of the IEEE international conference on computer vision, pp 1546–1554
- Pont-Tuset J, Arbelaez P, Barron JT, Marques F, Malik J (2016) Multiscale combinatorial grouping for image segmentation and object proposal generation. *IEEE Trans Pattern Anal Mach Intell* 39(1):128–140
- Qi X, Liu Z, Shi J, Zhao H, Jia J (2016) Augmented feedback in semantic segmentation under image level supervision. In: ECCV, pp 90–105
- Rajchl M, Lee MCH, Oktay O, Kamnitsas K, Passerat-Palmbach J, Bai W, Kainz B, Rueckert D (2016) Deepcut: object segmentation from bounding box annotations using convolutional neural networks. *IEEE Trans Med Imaging* 36:674–683
- Redondo-Cabrera C, Baptista-Ríos M, López-Sastre RJ (2018) Learning to exploit the prior network knowledge for weakly supervised semantic segmentation. *IEEE Transactions on Image Processing* 28:3649–3661
- Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, pp 91–99
- Rodriguez A, Laio A (2014) Clustering by fast search and find of density peaks. *Science* 344(6191):1492–1496
- Rother C, Kolmogorov V, Blake AJ (2004) “grabcut”: interactive foreground extraction using iterated graph cuts. *ACM Trans Graph* 23:309–314
- Roy A, Todorovic S (2017) Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 7282–7291
- Russakovsky O, Bearman AL, Ferrari V, Fei-Fei L (2015) What’s the point: semantic segmentation with point supervision. In: ECCV, pp 549–565
- Saito K, Watanabe K, Ushiku Y, Harada T (2018) Maximum classifier discrepancy for unsupervised domain adaptation. In: The IEEE conference on computer vision and pattern recognition (CVPR), pp 3723–3732
- Saleh F, Akbarian MSA, Salzmann M, Petersson L, Gould S, Alvarez JM (2016) Built-in foreground/background prior for weakly-supervised semantic segmentation. In: ECCV, pp 413–432
- Schroff F, Kalenichenko D, Philbin J (2015) Facenet: a unified embedding for face recognition and clustering. In: 2015 IEEE conference on computer vision and pattern recognition (CVPR), pp 815–823
- Scudder H (1965) Probability of error of some adaptive pattern-recognition machines. *IEEE Trans Inf Theory* 11(3):363–371
- Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2016) Grad-cam: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE international conference on computer vision (ICCV), pp 618–626
- Shaban M, Mahmood A, Al-Maadeed SA, Rajpoot N (2019) An information fusion framework for person localization via body pose in spectator crowds. *Inf Fusion* 51:178–188
- Shen D, Wu G, Suk HI (2017) Deep learning in medical image analysis. *Annu Rev Biomed Eng* 19(1):221–248
- Shen T, Lin G, Shen C, Reid ID (2018) Bootstrapping the performance of weakly supervised semantic segmentation. In: 2018 IEEE/CVF conference on computer vision and pattern recognition, pp 1363–1371
- Shi J, Malik J (2000) Normalized cuts and image segmentation. *IEEE Trans Pattern Anal Mach Intell* 22:888–905
- Shotton J, Johnson M, Cipolla R (2008) Semantic texton forests for image categorization and segmentation. In: 2008 IEEE conference on computer vision and pattern recognition, pp 1–8
- Siam M, Elkerdawy S, Jägersand M, Yogamani S (2017) Deep semantic segmentation for automated driving: taxonomy, roadmap and challenges. In: 2017 IEEE 20th international conference on intelligent transportation systems (ITSC), pp 1–8
- Singh KK, Lee YJ (2017) Hide-and-seek: forcing a network to be meticulous for weakly-supervised object and action localization. In: 2017 IEEE international conference on computer vision (ICCV), pp 3544–3553
- Souly N, Spampinato C, Shah M (2017) Semi supervised semantic segmentation using generative adversarial network. In: 2017 IEEE international conference on computer vision (ICCV), pp 5689–5697

- Sultana M, Mahmood A, Javed S, Jung SK (2019) Unsupervised deep context prediction for background estimation and foreground segmentation. *Mach Vis Appl* 30(3):375–395
- Tang M, Djelouah A, Perazzi F, Boykov Y, Schroers C (2018) Normalized cut loss for weakly-supervised cnn segmentation. In: 2018 IEEE/CVF conference on computer vision and pattern recognition, pp 1818–1827
- Thoma M (2016) A survey of semantic segmentation. [arXiv:1602.06541](https://arxiv.org/abs/1602.06541)
- Vernaza P, Chandraker MK (2017) Learning random-walk label propagation for weakly-supervised semantic segmentation. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 2953–2961
- Vezhnevets A, Buhmann JM (2010) Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning. In: 2010 IEEE computer society conference on computer vision and pattern recognition, pp 3249–3256
- Vezhnevets A, Ferrari V, Buhmann JM (2011) Weakly supervised semantic segmentation with a multi-image model. In: 2011 international conference on computer vision, pp 643–650
- Vezhnevets A, Ferrari V, Buhmann JM (2012) Weakly supervised structured output learning for semantic segmentation. In: 2012 IEEE conference on computer vision and pattern recognition, pp 845–852
- Wang H, Wang YI, Zhang Q, Xiang S, Pan C (2017) Gated convolutional neural network for semantic segmentation in high-resolution images. *Remote Sensing* 9:446
- Wang X, You S, Li X, Ma H (2018) Weakly-supervised semantic segmentation by iteratively mining common object features. In: 2018 IEEE/CVF conference on computer vision and pattern recognition, pp 1354–1362
- Wei Y, Xia W, Huang J, Ni B, Dong J, Zhao Y, Yan S (2014) Cnn: single-label to multi-label. *Computer Science* [arXiv:1406.5726](https://arxiv.org/abs/1406.5726)
- Wei Y, Liang X, Chen Y, Jie Z, Xiao Y, Zhao Y, Yan S (2016a) Learning to segment with image-level annotations. *Pattern Recognit* 59:234–244
- Wei Y, Liang X, Chen Y, Shen X, Cheng MM, Feng J, Zhao Y, Yan S (2016b) Stc: a simple to complex framework for weakly-supervised semantic segmentation. *IEEE Trans Pattern Anal Mach Intell* 39(11):2314–2320
- Wei Y, Feng J, Liang X, Cheng MM, Zhao Y, Yan S (2017) Object region mining with adversarial erasing: a simple classification to semantic segmentation approach. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 6488–6496
- 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: 2018 IEEE/CVF conference on computer vision and pattern recognition, pp 7268–7277
- Weston J, Ratle F, Mobahi H, Collobert R (2012) Deep learning via semi-supervised embedding. In: *Neural networks: tricks of the trade*, pp 639–655
- Wu Z, Han X, Lin YL, Gokhan Uzunbas M, Goldstein T, Nam Lim S, Davis LS (2018) Dcan: dual channel-wise alignment networks for unsupervised scene adaptation. In: *The European conference on computer vision (ECCV)*, pp 518–534
- Xu J, Schwing AG, Urtasun R (2015) Learning to segment under various forms of weak supervision. In: 2015 IEEE conference on computer vision and pattern recognition (CVPR), pp 3781–3790
- Yu H, Yang Z, Tan L, Wang Y, Sun W, Sun M, Tang Y (2018) Methods and datasets on semantic segmentation: a review. *Neurocomputing* 304:82–103
- Zhan X, Liu Z, Luo P, Tang X, Loy CC (2017) Mix-and-match tuning for self-supervised semantic segmentation. In: *AAAI*
- Zhang H, Fritts JE, Goldman SA (2008) Image segmentation evaluation: A survey of unsupervised methods. *computer vision and image understanding* 110(2):260–280
- Zhang L, Yang Y, Gao Y, Yu Y, Wang C, Li X (2014) A probabilistic associative model for segmenting weakly supervised images. *IEEE Transactions on Image Processing* 23:4150–4159
- Zhang W, Huang H, Schmitz M, Sun X, Wang H, Mayer H (2017) Effective fusion of multi-modal remote sensing data in a fully convolutional network for semantic labeling. *Remote Sens* 10:52
- Zhang X, Wei Y, Feng J, Yang Y, Huang TS (2018a) Adversarial complementary learning for weakly supervised object localization. In: 2018 IEEE/CVF conference on computer vision and pattern recognition, pp 1325–1334
- Zhang Y, Qiu Z, Yao T, Liu D, Mei T (2018b) Fully convolutional adaptation networks for semantic segmentation. In: *The IEEE conference on computer vision and pattern recognition (CVPR)*, pp 6810–6818
- Zhang SH, Li R, Dong X, Rosin P, Cai Z, Han X, Yang D, Huang H, Hu SM (2019) Pose2seg: detection free human instance segmentation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 889–898

- Zhao B, Wang F, Zhang C (2008) Efficient multiclass maximum margin clustering. In: ICML, pp 1248–1255
- Zhao B, Kwok JT, Zhang C (2009) Maximum margin clustering with multivariate loss function. In: 2009 ninth IEEE international conference on data mining, pp 637–646
- Zhao A, Balakrishnan G, Durand F, Guttag JV, Dalca AV (2019) Data augmentation using learned transformations for one-shot medical image segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8543–8553
- Zhou B, Khosla A, Lapedriza À, Oliva A, Torralba A (2015) Learning deep features for discriminative localization. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR), pp 2921–2929
- Zhou Y, Zhu Y, Ye Q, Qiu Q, Jiao J (2018) Weakly supervised instance segmentation using class peak response. In: 2018 IEEE/CVF conference on computer vision and pattern recognition, pp 3791–3800
- Zhou Y, Liu X, Zhao J, Ma D, Yao R, Liu B (2019) Zheng Y (2019) Remote sensing scene classification based on rotation-invariant feature learning and joint decision making. EURASIP J Image Video Process 1:3

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.