A Survey on Deep Learning-based Architectures for Semantic Segmentation on 2D Images

Irem Ulku & Erdem Akagündüz

Published online: 09 Feb 2022.

Submit your article to this journal ✐

Article views: 1445

View related articles ✐

View Crossmark data ✐

A Survey on Deep Learning-based Architectures for Semantic Segmentation on 2D Images

Irem Ulku [ID][a] and Erdem Akagündüz [ID][b]

[a]Department of Computer Engineering, Ankara University, Ankara, Turkey; [b]Graduate School of Informatics Middle East Technical University, Ankara, Turkey

ABSTRACT

Semantic segmentation is the pixel-wise labeling of an image. Boosted by the extraordinary ability of convolutional neural networks (CNN) in creating semantic, high-level and hierarchical image features; several deep learning-based 2D semantic segmentation approaches have been proposed within the last decade. In this survey, we mainly focus on the recent scientific developments in semantic segmentation, specifically on deep learning-based methods using 2D images. We started with an analysis of the public image sets and leaderboards for 2D semantic segmentation, with an overview of the techniques employed in performance evaluation. In examining the evolution of the field, we chronologically categorized the approaches into three main periods, namely pre- and early deep learning era, the fully convolutional era, and the post-FCN era. We technically analyzed the solutions put forward in terms of solving the fundamental problems of the field, such as fine-grained localization and scale invariance. Before drawing our conclusions, we present a table of methods from all mentioned eras, with a summary of each approach that explains their contribution to the field. We conclude the survey by discussing the current challenges of the field and to what extent they have been solved.

Introduction

Semantic segmentation has recently become one of the fundamental problems, and accordingly, a hot topic for the fields of computer vision and machine learning. Assigning a separate class label to each pixel of an image is one of the important steps in building complex robotic systems, such as driverless cars/drones, human-friendly robots, robot-assisted surgery, and intelligent military systems. Thus, it is no wonder that in addition to scientific institutions, industry-leading companies studying artificial intelligence are now summarily confronting this problem.

The simplest problem definition for semantic segmentation is pixel-wise labeling. Because the problem is defined at the pixel level, finding only class labels that the scene includes is considered insufficient, but localizing labels at the original image pixel resolution is also a fundamental goal. Depending on the context, class

CONTACT Erdem Akagündüz ✉ akaerdem@metu.edu.tr 🖃 Graduate School of Informatics Middle East Technical University, Turkey

labels may change. For example, in a driverless car, the pixel labels may be *human, road*, and *car* (Siam et al. 2017), whereas for a medical system (Jiang et al. 2017; Saha and Chakraborty 2018), they could be *cancer cells, muscle tissue, aorta wall* etc.

The recent increase in interest in this topic has been undeniably caused by the extraordinary success seen with convolutional neural networks (LeCun et al. 1989) (CNN) that have been brought to semantic segmentation. Understanding a scene at the semantic level has long been one of the main topics of computer vision, but it is only now that we have seen actual solutions to the problem.

In this paper, our primary motivation is to focus on the recent scientific developments in semantic segmentation, specifically on the evolution of deep learning-based methods using 2D images. The reason we narrowed down our survey to techniques that utilize only 2D visible imagery is that, in our opinion, the scale of the problem in the literature is so vast and widespread that it would be impractical to analyze and categorize all semantic segmentation modalities (such as 3D point clouds, hyper-spectral data, MRI, CT,[1] etc.) found in journal articles to any degree of detail. In addition to analyzing, the techniques that make semantic segmentation possible and accurate, we also examine the most popular image sets created for this problem. Additionally, we review the performance measures used for evaluating the success of semantic segmentation. Most importantly, we propose a taxonomy of methods, together with a technical evolution of them, which we believe is novel in the sense that it provides insight to the existing deficiencies and suggests future directions for the field.

The remainder of the paper is organized as follows: in the following subsection, we refer to other survey studies on the subject and underline our contribution. Section 2 presents information about the different image sets, the challenges, and how to measure the performance of semantic segmentation. Starting with Section 3, we chronologically scrutinize semantic segmentation methods under three main titles, hence in three separate sections. Section 3 covers the methods of pre- and early deep convolutional neural networks era. Section 4 provides details on the fully convolutional neural networks, which we consider a milestone for the semantic segmentation literature. Section 5 covers the state-of-the-art methods on the problem and provides details on both the architectural details and the success of these methods. Before finally concluding the paper in Section 7, Section 6 provides a future scope and potential directions for the field.

Surveys on Semantic Segmentation

Very recently, driven by both academia and industry, the rapid increase of interest in semantic segmentation has inevitably led to a number of survey studies being published (Thoma 2016; Ahmad et al. 2017; Jiang et al. 2017; Siam et al. 2017; Garcia-Garcia et al. 2017; Saffar et al. 2018; Yu et al. 2018; Guo et al. 2018; Lateef and Ruichek 2019; Minaee et al. 2020).

Some of these surveys focus on a specific problem, such as comparing semantic segmentation approaches for horizon/skyline detection (Ahmad et al. 2017), whilst others deal with relatively broader problems related to industrial challenges, such as semantic segmentation for driverless cars (Siam et al. 2017) or medical systems (Jiang et al. 2017). These studies are useful if working on the same specific problem, but they lack an overarching vision that may 'technically' contribute to the future directions of the field.

Another group (Thoma 2016; Saffar et al. 2018; Yu et al. 2018; Guo et al. 2018) of survey studies on semantic segmentation have provided a general overview of the subject, but they lack the necessary depth of analysis regarding deep learning-based methods. Whilst semantic segmentation was studied for two decades prior to deep learning, actual contribution to the field has only been achieved very recently, particularly following a revolutionary paper on fully convolutional networks (FCN) (Shelhamer, Long, and Darrell 2017) (which has also been thoroughly analyzed in this paper). It could be said that most state-of-the-art studies are in fact extensions of that same (Shelhamer, Long, and Darrell 2017) study. For this reason, without scrupulous analysis of FCNs and the direction of the subsequent papers, survey studies will lack the necessary academic rigor in examining semantic segmentation using deep learning.

There are recent reviews of deep semantic segmentation, for example, by (Garcia-Garcia et al. 2017) and (Minaee et al. 2020), which provide a comprehensive survey on the subject. These survey studies cover almost all the popular semantic segmentation image sets and methods, and for all modalities, such as 2D, RGB, 2.5D, RGB-D, and 3D data. Although they are inclusive in the sense that most related material on deep semantic segmentation is included, the categorization of the methods is coarse, since the surveys attempt to cover almost everything umbrellaed under the topic of semantic segmentation literature.

A detailed categorization of the subject was provided in (Lateef and Ruichek 2019). Although this survey provides important details on the subcategories that cover almost all approaches in the field, discussions on how the proposed techniques are chronologically correlated are left out of their scope. Recent deep learning studies on semantic segmentation follow a number of fundamental directions and labor with tackling the varied corresponding issues. In this survey paper, we define and describe these new challenges, and present the chronological evolution of the techniques of all the studies within this proposed context. We believe our attempt to understand the evolution of semantic segmentation architectures is the main contribution of the paper. We provide a table of these related methods, and explain them briefly one after another in chronological order, with their metric performance and

computational efficiency. This way, we believe that readers will better understand the evolution, current state-of-the-art, as well as the future directions seen for 2D semantic segmentation.

Image Sets, Challenges, and Performance Evaluation

Image Sets and Challenges

The level of success for any machine-learning application is undoubtedly determined by the quality and the depth of the data being used for training. When it comes to deep learning, data is even more important since most systems are termed end-to-end; thus, even the features are determined *by* the data, not *for* the data. Therefore, data is no longer the object but becomes the actual subject in the case of deep learning.

In this section, we scrutinize the most popular large-scale 2D image sets that have been utilized for the semantic segmentation problem. The image sets were categorized into two main branches, namely general-purpose image sets, with generic class labels including almost every type of object or background, and also urban street image sets, which include class labels such as car and person, and are generally created for the training of driverless car systems. There are many other unresolved 2D semantic segmentation problem domains, such as medical imaging, satellite imagery, or infrared imagery. However, urban street image is currently driving scientific development in the field because they attract more attention from the industry. Therefore, very large-scale image sets and challenges with crowded leaderboards exist, yet, only specifically for industrial users. Scientific interest for depth-based semantic segmentation is growing rapidly; however, as mentioned in the Introduction, we have excluded depth-based and 3D-based segmentation datasets from the current study in order to focus with sufficient detail on the novel categorization of recent techniques pertinent to 2D semantic segmentation.

General Purpose Semantic Segmentation Image Sets

• PASCAL Visual Object Classes (VOC) (Everingham et al. 2010): This image set includes image annotations not only for semantic segmentation, but for also classification, detection, action classification, and person layout tasks. The image set and annotations are regularly updated and the leaderboard of the challenge is public[2] (with more than 140 submissions just for the segmentation challenge alone). It is the most popular among the semantic segmentation challenges and is still active following its initial release in 2005. The PASCAL VOC semantic segmentation challenge image set includes 20 foreground object classes and one background class. The original data consisted of 1,464

images for the purposes of training, plus 1,449 images for validation. The 1,456 test images are kept private for the challenge. The image set includes all types of indoor and outdoor images and is generic across all categories.

The PASCAL VOC image set has a number of extension image sets, the most popular among these are PASCAL Context (Mottaghi et al. 2014) and PASCAL Parts (Chen et al. 2014b). The first (Mottaghi et al. 2014) is a set of additional annotations for PASCAL VOC 2010, which goes beyond the original PASCAL semantic segmentation task by providing annotations for the whole scene. The statistics section contains a full list of more than 400 labels (compared to the original 21 labels). The second (Chen e tal. 2014b) is also a set of additional annotations for PASCAL VOC 2010. It provides segmentation masks for each body part of the object, such as the separately labeled limbs and body of an animal. For these extensions, the training and validation set contains 10,103 images, while the test set contains 9,637 images. There are other extensions to PASCAL VOC using other functional annotations such as the Semantic Parts (PASParts) (Wang et al. 2015) image set and the Semantic Boundaries Dataset (SBD) (Hariharan et al. 2011). For example, PASParts (Wang et al. 2015) additionally provides 'instance' labels, such as two instances of an object within an image are labeled separately, rather than using a single class label. However, unlike the former two additional extensions (Chen et al. 2014b; Mottaghi et al. 2014), these further extensions (Hariharan et al. 2011; Wang et al. 2015) have proven less popular as their challenges have attracted much less attention in state-of-the-art semantic segmentation studies; thus, their leaderboards are less crowded. In Figure 1, a sample object, parts and instance segmentation are depicted.

• Common Objects in Context (COCO) (Lin et al. 2014): With 200 K labeled images, 1.5 million object instances, and 80 object categories, COCO is a very large-scale object detection, semantic segmentation, and captioning image set, including almost every possible types of scene. COCO provides challenges not only at the instance-level and pixel-level (which they refer to as *stuff*) semantic segmentation, but also introduces a novel task, namely that of *panoptic* segmentation (Kirillov et al. 2018), which aims at unifying instance-level and pixel-level segmentation tasks. Their leaderboards[3] are relatively less crowded because of the scale of the data. On the other hand, for the same reason, their challenges are assessed only by the most ambitious scientific and
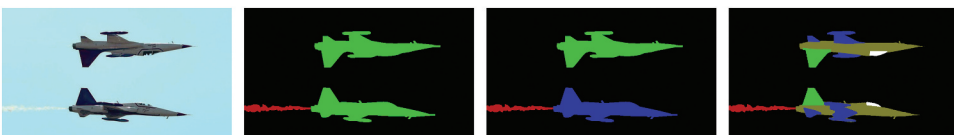

Figure 1. A sample image and its annotation for object, instance and parts segmentations separately, from left to right.

industrial groups, and thus are considered as the state-of-the-art in their leaderboards. Due to its extensive volume, most studies partially use this image set to pre-train or fine-tune their model, before submitting to other challenges such as PASCAL VOC 2012.

• ADE20K dataset (Zhou et al. 2019): ADE20K contains more than 20 K scene-centric images with objects and object parts annotations. Similarly to PASCAL VOC, there is a public leaderboard[4] and the benchmark is divided into 20 K images for training, 2 K images for validation, and another batch of held-out images for testing. The samples in the dataset have varying resolutions (average image size being 1.3 M pixels), which can be up to 2400 × 1800 pixels. There are total of 150 semantic categories included for evaluation.

• Other General Purpose Semantic Segmentation Image Sets: Although less popular than either PASCAL VOC or COCO, there are also some other image sets in the same domain. Introduced by (Prest et al. 2012), YouTube-Objects is a set of low-resolution (480 × 360) video clips with more than 10k pixel-wise annotated frames.

Similarly, SIFT-flow (Tighe and Lazebnik 2010) is another low-resolution (256 × 256) semantic segmentation image set with 33 class labels for a total of 2,688 images. These and other relatively primitive image sets have been mostly abandoned in the semantic segmentation literature due to their limited resolution and low volume.

Urban Street Semantic Segmentation Image Sets

• Cityscapes (Cordts et al. 2016): This is a large-scale image set with a focus on the semantic understanding of urban street scenes. It contains annotations for high-resolution images from 50 different cities, taken at different hours of the day and from all seasons of the year, and also with varying backgrounds and scene layouts. The annotations are carried out at two quality levels: fine for 5,000 images and course for 20,000 images. There are 30 different class labels, some of which also have instance annotations (vehicles, people, riders, etc.). Consequently, there are two challenges with separate public leaderboards[5]: one for pixel-level semantic segmentation, and a second for instance-level semantic segmentation. There are more than 100 entries to the challenge, making it the most popular regarding semantic segmentation of urban street scenes.

• Other Urban Street Semantic Segmentation Image Sets: There are a number of alternative image sets for urban street semantic segmentation, such as Cam-Vid (Brostow, Fauqueur, and Cipolla 2009), KITTI (Geiger et al. 2013), SYNTHIA (Ros et al. 2016a), and IDD (Varma et al. 2018). These are generally overshadowed by the Cityscapes image set (Cordts et al. 2016) for several reasons. Principally, their scale is relatively low. Only the SYNTHIA image set (Ros et al. 2016a) can be considered as large scale (with more than

13k annotated images); however, it is an artificially generated image set, and this is considered a major limitation for security-critical systems like driverless cars.

Small-scale and Imbalanced Image Sets

In addition to the aforementioned large-scale image sets of different categories, there are several image sets with insufficient scale or strong imbalance such that, when applied to deep learning-based semantic segmentation models, high-level segmentation accuracies cannot be directly obtained. Most public challenges on semantic segmentation include sets of this nature such as the DSTL or RIT-18 (DSTLab; Kemker, Salvaggio, and Kanan 2018), just to name a few. Because of the overwhelming numbers of these types of sets, we chose to include only the details of the large-scale sets that attract the utmost attention from the field.

Nonetheless, being able to train a model that performs well on small-scale or imbalanced data is a correlated problem to ours. Besides conventional deep learning techniques, such as transfer learning or data augmentation; the problem of insufficient or imbalanced data can be attacked by using specially designed deep learning architectures such as some optimized convolution layer types (Chen et al. (2018a); He et al. (2015), etc.) and others that we cover in this survey paper. What is more, there are recent studies that focus on the specific problem of utilizing insufficient sets for the problem of deep learning-based semantic segmentation (Xia, Lu, and Gu 2019). Although we acknowledge this problem as fundamental for the semantic segmentation field, we leave the discussions on techniques to handle small-scale or imbalanced sets for semantic segmentation, beyond the scope of this survey paper.

Performance Evaluation

There are two main criteria in evaluating the performance of semantics segmentation: accuracy, or in other words, the success of an algorithm; and computation complexity in terms of speed and memory requirements. In this section, we analyze these two criteria separately.

Accuracy

Measuring the performance of segmentation can be complicated, mainly because there are two distinct values to measure. The first is classification, which is simply determining the pixel-wise class labels; and the second is localization, or finding the correct set of pixels that enclose the object. Different metrics can be found in the literature to measure one or both of these values. The following is a brief explanation of the principal measures most commonly used in evaluating semantic segmentation performance.

• *ROC-AUC*: ROC stands for the Receiver-Operator Characteristic curve, which summarises the trade-off between true positive rate and false-positive rate for a predictive model using different probability thresholds; whereas AUC stands for the area under this curve, which is 1 at maximum. This tool is useful in interpreting binary classification problems and is appropriate when observations are balanced between classes. However, since most semantic segmentation image sets (Everingham et al. 2010; Mottaghi et al. 2014; Chen et al. 2014b; Wang et al. 2015; Hariharan et al. 2011; Lin et al. 2014; Cordts et al. 2016) are not balanced between the classes, this metric is no longer used by the most popular challenges.

• *Pixel Accuracy*: Also known as *global accuracy* (Badrinarayanan, Kendall, and Cipolla 2015), pixel accuracy (PA) is a very simple metric which calculates the ratio between the amount of properly classified pixels and their total number. Mean pixel accuracy (mPA), is a version of this metric which computes the ratio of correct pixels on a per-class basis. mPA is also referred to as *class average accuracy* (Badrinarayanan, Kendall, and Cipolla 2015).

$$PA = \frac{\sum_{j=1}^{k} n_{jj}}{\sum_{j=1}^{k} t_j}, \qquad mPA = \frac{1}{k} \sum_{j=1}^{k} \frac{n_{jj}}{t_j} \tag{1}$$

where n_{jj} is the total number of pixels both classified and labeled as class *j*. In other words, n_{jj} corresponds to the total number of *True Positives* for class *j*. t_j is the total number of pixels labeled as class *j*.

• *Intersection over Union* (IoU): Also known as the Jaccard Index, IoU is a statistic used for comparing the similarity and diversity of sample sets. In semantics segmentation, it is the ratio of the intersection of the pixel-wise classification results with the ground truth, to their union.

$$IoU = \frac{\sum_{j=1}^{k} n_{jj}}{\sum_{j=1}^{k} \left(n_{ij} + n_{ji} + n_{jj} \right)}, \qquad i \neq j \tag{2}$$

where, n_{ij} is the number of pixels, which are labeled as class *i*, but classified as class *j*. In other words, they are *False Positives* (false alarms) for class *j*. Similarly, n_{ji}, the total number of pixels labeled as class *j*, but classified as class *i* are the *False Negatives* (misses) for class *j*.

Two extended versions of IoU are also widely in use:

• *Mean Intersection over Union* (mIoU): mIoU is the class-averaged IoU, as in (3).

$$mIoU = \frac{1}{k} \sum_{j=1}^{k} \frac{n_{jj}}{n_{ij} + n_{ji} + n_{jj}}, \qquad i \neq j \tag{3}$$

○ *Frequency-weighted IoU* (FwIoU): This is an improved version of MIoU that weighs each class importance depending on appearance frequency by using t_j (the total number of pixels labeled as class j, as also defined in (1)). The formula of FwIoU is given in (4):

$$FwIoU = \frac{1}{\sum_{j=1}^{k} t_j} \sum_{j=1}^{k} t_j \frac{n_{jj}}{n_{ij} + n_{ji} + n_{jj}}, \qquad i \neq j \qquad (4)$$

IoU and its extensions, compute the ratio of true positives (hits) to the sum of false positives (false alarms), false negatives (misses) and true positives (hits). Thereby, the IoU measure is more informative when compared to pixel accuracy simply because it takes false alarms into consideration, whereas PA does not. However, since false alarms and misses are summed up in the denominator, the significance between them is not measured by this metric, which is considered its primary drawback. In addition, IoU only measures the number of pixels correctly labeled without considering how accurate the segmentation boundaries are.

• *Precision-Recall Curve (PRC)-based metrics*: Precision (ratio of hits over a summation of hits and false alarms) and recall (ratio of hits over a summation of hits and misses) are the two axes of the PRC used to depict the trade-off between precision and recall, under a varying threshold for the task of binary classification. PRC is very similar to ROC. However, PRC is more powerful in discriminating the effects between the false positives (alarms) and false negatives (misses). That is predominantly why PRC-based metrics are commonly used for evaluating the performance of semantic segmentation. The formula for Precision (also called Specificity) and Recall (also called Sensitivity) for a given class j, are provided in (5):

$$Prec. = \frac{n_{jj}}{n_{ij} + n_{jj}}, \quad Recall = \frac{n_{jj}}{n_{ji} + n_{jj}}, i \neq j \qquad (5)$$

There are three main PRC-based metrics:

○ F_{score} : Also known as the '*dice coefficient*,' this measure is the harmonic mean of the precision and recall for a given threshold. It is a normalised measure of similarity, and ranges between 0 and 1 (Please see (6)).

$$F_{score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad (6)$$

○ *PRC-AuC*: This is similar to the ROC-AUC metric. It is simply the area under the PRC. This metric refers to information about the precision-recall trade-off for different thresholds, but not the *shape* of the PR curve.

- *Average Precision* (AP): This metric is a single value that summarizes both the shape and the AUC of PRC. In order to calculate AP, using the PRC, for uniformly sampled recall values (e.g., 0.0, 0.1, 0.2, . . ., 1.0), precision values are recorded. The average of these precision values is referred to as the average precision. This is the most commonly used single value metric for semantic segmentation. Similarly, mean average precision (mAP) is the mean of the AP values, calculated on a per-class basis.

• *Hausdorff Distance* (HD): Hausdorff Distance is used incorporating the longest distance between classified and labeled pixels as an indicator of the largest segmentation error (Jadon 2020; Karimi and Salcudean 2019), with the aim of tracking the performance of a semantic segmentation model. The unidirectional HDs as $hd(X, Y)$ and $hd(Y, X)$ are presented in (7) and (8), respectively.

$$hd(X, Y) = \max_{x \varepsilon X} \min_{y \varepsilon Y} \|x - y\|_2, \tag{7}$$

$$hd(Y, X) = \max_{y \varepsilon Y} \min_{x \varepsilon X} \|x - y\|_2. \tag{8}$$

where, X and Y are the pixel sets. The x is the pixel in the segmented counter X and y is the pixel in the target counter Y (Huang et al. 2020). The bidirectional HD between these sets is shown in (9), where the Euclidean distance is employed for (7), (8), and (9).

$$HD(X, Y) = max(hd(X, Y), hd(Y, X)). \tag{9}$$

IoU and its variants, along with AP, are the most commonly used accuracy evaluation metrics in the most popular semantic segmentation challenges (Everingham et al. 2010; Mottaghi et al. 2014; Chen et al. 2014b; Wang et al. 2015; Hariharan et al. 2011; Lin et al. 2014; Cordts et al. 2016).

Computational Complexity

The burden of computation is evaluated using two main metrics: how fast the algorithm completes, and how much computational memory is demanded.

- *Execution time*: This is measured as the whole processing time, starting from the instant a single image is introduced to the system/algorithm right through until the pixel-wise semantic segmentation results are obtained. The performance of this metric significantly depends on the hardware utilized. Thus, for an algorithm, any execution time metric should be accompanied by a thorough description of the hardware used. There are notations such as Big-O, which provide a complexity measure independent of the implementation domain. However, these notations are highly theoretical and are predominantly not preferred for extremely complex algorithms such as deep semantic segmentation as they are simple and largely inaccurate.

For a deep learning-based algorithm, the offline (i.e., training) and online (i.e., testing) operation may last for considerably different time intervals. Technically, the execution time refers only to the online operation or academically speaking, the test duration for a single image. Although this metric is extremely important for industrial applications, academic studies refrain from publishing exact execution times, and none of the aforementioned challenges was found to have provided this metric. A recent study, (Zhao et al. 2018a) provided a 2D histogram of Accuracy (MIoU%) vs frames-per-second, in which some of the state-of-the-art methods with open-source codes (including their proposed structure, namely image cascade network – ICNet), were benchmarked using the Cityscapes (Cordts et al. 2016) image set.

• *Memory Usage*: Memory usage is specifically important when semantic segmentation is utilized in limited performance devices, such as smartphones, digital cameras, or when the requirements of the system are extremely restrictive. The prime examples of these would be military systems or security-critical systems, such as self-driving cars.

The usage of memory for a complex algorithm like semantic segmentation may change drastically during operation. That is why a common metric for this purpose is *peak memory usage*, which is simply the maximum memory required for the entire segmentation operation for a single image. The metric may apply to computer (data) memory or GPU memory depending on the hardware design.

Although critical for industrial applications, this metric is not usually made available for any of the aforementioned challenges.

Computational efficiency is a very important aspect of any algorithm that is to be implemented on a real system. A comparative assessment of the speed and capacity of various semantic segmentation algorithms is a challenging task. Although most state-of-the-art algorithms are available with open-source codes, benchmarking all of them, with their optimal hyper-parameters, seems implausible. For this purpose, we provide an inductive way of comparing the computational efficiencies of methods in the following sections. In Table 1, we categorize methods into mainly four levels of computational efficiency and discuss our categorization related to the architectural design of a given method. This table also provides a chronological evolution of the semantic segmentation methods in the literature.

Before Fully Convolutional Networks

As mentioned in the Introduction, the utilization of FCNs is a breaking point for semantic segmentation literature. Efforts on semantic segmentation literature prior to FCNs (Shelhamer, Long, and Darrell 2017) can be analyzed in two separate branches, as pre-deep learning and early deep learning approaches. In this section, we briefly discuss both sets of approaches.

Table 1. State-of-the-art semantic segmentation methods, showing the method name and reference, brief summary, problem type targeted, and refinement model (if any).

Method	Method Summary	Rankings	Eff.
MultiScale-Net. (Farabet et al. 2013)	*Multiscale convolutional network fused parallel with a segmentation framework (either superpixel or CRF-based). Relatively lower computational efficiency due to a CRF block.*	68.7% mPA @SIFTflow	★ ★
Recurrent CNN (Pinheiro and Collobert 2014)	*Recurrent architecture constructed by using different instances of a CNN, in which each network instance is fed with previous label predictions (obtained from the previous instance). Heavy computational load when multiple instances (3 in their best performing experiments) are fed.*	77.7% mPA @SIFTflow	★
FCN (Shelhamer, Long, and Darrell 2017)	*Fully convolutional encoder structure (i.e., no fully connected layers) with skip connections that fuse multiscale activations at the final decision layer. Relative fast due to no fully connected layers or a refinement block.*	85.2% mPA @SIFTflow 62.2% mIoU @PASCAL 2012 65.3% mIoU @CitySca. (w/o course) 39.3% mIoU @ADE20K	★ ★ ★
DeepLab.v1 (Chen et al. 2014a)	*CNN with dilated convolutions, succeeded by a fully-connected (i.e. Dense) CRF. Fast and optimized computation leads to near real-time performance.*	66.4% mIoU @PASCAL 2012	★ ★ ★
CMSA (Eigen and Fergus 2014)	*Layers of a pyramidal input are fed to separate FCNs for different scales in parallel. These multiscale FCNs are also connected in series to provide pixel-wise category, depth and normal output, simultaneously. Relatively lower computational efficiency due to progressive processing of sequence of different scales.*	83.8% mPA @SIFTflow 62.6% mIoU @PASCAL 2012	★ ★
UNet (Ronneberger, Fischer, and Brox 2015)	*Encoder/decoder structure with skip connections that connect same levels of ED and final input-sized classification layer. Efficient computation load due to no fully connected layers or a refinement block.*	72.7% mIoU @PASCAL 2012 (tested by (Zhang et al. 2018c))	★ ★ ★
SegNet (Badrinarayanan, Kendall, and Cipolla 2015)	*Encoder/decoder structure (similar to UNet) with skip connections that transmit only pooling indices (unlike U-Net, for which skip connections concatenate same-level activations. Efficient computation load due to no fully connected layers or a refinement block).*	59.9% mIoU @PASCAL 2012 79.2% mIoU @CitySca. (w/o course)	★ ★ ★
DeconvNet (Noh, Hong, and Han 2015)	*Encoder/decoder structure (namely 'the Conv./ Deconv. Network') without skip connections. The encoder (convolutional) part of the network is transferred from the VGG-VD-16 L. Efficient computation load due to no fully connected layers or a refinement block. (Simonyan and Zisserman 2015).*	74.8% mIoU @PASCAL 2012	★ ★ ★
MSCG (Yu and Koltun 2015)	*Multiscale context aggregation using only a rectangular prism of dilated convolutional layers, without pooling or subsampling layers, to perform pixel-wise labelling. Efficient computation load due to no fully connected layers or a refinement block.*	67.6% mIoU @PASCAL 2012 67.1% mIoU @CitySca. (w/o course)	★ ★ ★
CRF-as-RNN (Zheng et al. 2015)	*Fully convolutional CNN (i.e., FCN) followed by a CRF-as-RNN layer, in which an iterative CRF algorithm is formulated as an RNN. Because of the RNN block, computational efficiency is limited.*	65.2% mIoU @PASCAL 2012 62.5% mIoU @CitySca. (w/o course)	★ ★

(*Continued*)

Table 1. (Continued).

Method	Method Summary	Rankings	Eff.
FeatMap-Net. (Lin et al. 2016a)	Layers of a pyramidal input fed to parallel multiscale feature maps (i.e., CNNS), and later fused in an upsample/concatenation (i.e. pyramid pooling) layer to provide the final feature map to a Dense CRF Layer. Well-planned but loaded architecture leads to moderate computational efficiency.	88.1% mPA @SIFTflow 75.3% mIoU @PASCAL 2012	★ ★
Graph LSTM (Liang et al. 2016)	Generalization of LSTM from sequential data to general graph-structured data for semantic segmentation on 2D still images, mostly people/parts. Graph-LSTM processing considerably limits computation efficiency.	60.2% mIoU @PASCAL Person/Parts 2010	★
DAG-RNN (Shuai et al. 2016)	DAG-structured CNN+RNN network that models long-range semantic dependencies among image units. Due to chain structured sequential processing of pixels with a recurrent model, the computational efficiency is considerably limited.	85.3% mPA @SIFTflow	★
DeepLab.v2 (Chen et al. 2018a)	Improved version of DeepLab.v1, with additional 'dilated (atrous) spatial pyramid pooling' (ASPP) layer. Similar computational performance to DeepLab.v1.	79.7% mIoU @PASCAL 2012 70.4% mIoU @CitySca. (w/o course)	★ ★ ★
PSPNet (Zhao et al. 2017)	CNN followed by a pyramid pooling layer similar to (He et al. 2015), but without a fully connected decision layer. Hence, computational performance closer to FCN (Shelhamer, Long, and Darrell 2017).	85.5% mIoU @PASCAL 2012 81.2% mIoU @CitySca. (w. course) 55.4% mIoU @ADE20K	★ ★ ★
DeepLab.v3 (Chen et al. 2017)	Improved version of DeepLab.v2, with optimization of ASPP layer hyperparameters and without a Dense CRF layer, for faster operation.	85.7% mIoU @PASCAL 2012 81.3% mIoU @CitySca. (w. course)	★ ★ ★
DIS (Luo et al. 2017)	One network predicts labelmaps/tags, while another performs semantic segmentation using these predictions. Both networks use ResNet101 (He et al. 2016) for preliminary feature extraction. They declare similar computational efficiency to DeepLabv2 (Chen et al. 2018a)	41.7% mIoU @COCO 86.8% mIoU @PASCAL 2012	★ ★ ★
Mask-RCNN (He et al. 2017)	Object Detector Fast-RCNN followed by ROI-pooling and Convolutional layers, applied to instance segmentation, with near real-time performance (see Figure 5.a).	37.1% mIoU @COCO tested by YOLACT2019	★ ★ ★
GCN (Peng et al. 2017)	Fed by an initial ResNet-based (He et al. 2016) encoder, GCN uses large kernels to fuse high- and low-level features in a multiscale manner, followed by a convolutional Border Refinement (BR) module. Its fully convolutional architectue allows near real-time performance.	83.6% mIoU @PASCAL 2012 76.9% mIoU @CitySca. (w/o course)	★ ★ ★
SDN (Fu et al. 2017)	UNET architecture that consists of multiple shallow deconvolutional networks, called SDN units, stacked one by one to integrate contextual information and guarantee fine recovery of localized information. Computational efficiency similar to UNET-like architectures.	83.5% mIoU @PASCAL 2012	★ ★ ★
DFN (Yu et al. 2018b)	Consists of two sub-networks: Smooth Net (SN) and Border Net (BN). SN utilizes an attention module and handles global context, whereas BN employs a refinement block to handle borders. Limited computational efficiency due to an attention block.	86.2% mIoU @PASCAL 2012 80.3% mIoU @CitySca. (w.course)	★ ★

(Continued)

Table 1. (Continued).

Method	Method Summary	Rankings	Eff.
MSCI (Lin et al. 2018)	*Aggregates features from different scales via connections between Long Short-term Memory (LSTM) chains. Limited computational efficiency due to multiple RNN blocks (i.e. LSTMs).*	88.0% mIoU @PASCAL 2012	★ ★
DeepLab.v3+ (Chen et al. 2018b)	*Improved version of DeepLab.v3, using special encoder-decoder structure with dilated convolutions (with no Dense CRF employed for faster operation).*	87.3% mIoU @PASCAL 2012 82.1% mIoU @CitySca. (w. course)	★ ★ ★
HPN (Shi et al. 2018)	*Followed by a convolutional 'Appearance Feature Encoder,' a 'Contextual Feature Encoder' consisting of LSTMs generates super-pixel features fed to a Softmax-based classification layer. Limited computational efficiency due to multiple LSTMs.*	85.8% mIoU @PASCAL 2012 92.3% mPA @SIFTflow	★ ★
EncNet (Zhang et al. 2018a)	*Fully connected structure to extract context is fed by dense feature maps (obtained from ResNet (He et al. 2016)) and followed by a convolutional prediction layer. Fully connected layers within their "Context Encoding Module" limits computational performance.*	85.9% mIoU @PASCAL 2012 55.7% mIoU @ADE20K	★ ★
PSANet (Zhao et al. 2018b)	*A convolutional point-wise spatial attention (PSA) module is attached to o pretrained convolutional encoder, so that pixels are interconnected through a self-adaptively learnt attention map to provide global context. Additional PSA module limits computational efficieny compared to fully convolutional architectures (e.g. FCN).*	85.7% mIoU @PASCAL 2012 81.4% mIoU @CitySca. (w. course)	★ ★
PAN (Li et al. 2018)	*SPP layer with global pooling architecture. Similar architecture and thus, computational efficiency with PSPNet (Zhao et al. 2017).*	84.0% mIoU @PASCAL 2012 (*taken from the paper, not listed in the leaderboard*)	★ ★ ★
ExFuse (Zhang et al. 2018c)	*Improved version of GCN (Peng et al. 2017) for feature fusing which introduces more semantic information into low-level features and more spatial details into high-level features, by additional skip connections. Computational performance comparable to GCN.*	87.9% mIoU @PASCAL 2012	★ ★ ★
EMANet152 (Li et al. 2019b)	*Novel attention module between two CNN structures converts input feature maps to output feature maps, thus providing global context. Computationally more efficient compared to other attention governing architectures (e.g. PSANet).*	88.2% mIoU @PASCAL 2012 39.9% mIoU @COCO	★ ★ ★
KSAC (Huang et al. 2019)	*Allows branches of different receptive fields to share the same kernel to facilitate communication among branches and perform feature augmentation inside the network. The idea is similar to ASPP layer of DeepLabv3 (Chen et al. 2017), hence similar computational performance.*	88.1% mIoU @PASCAL 2012	★ ★ ★
CFNet (Zhang et al. 2019)	*Using a distribution of co-occurrent features for a given target in an image, a fine-grained spatial invariant representation is learnt and the CFNet is constructed. Similar architecture to PSANet (Zhao et al. 2018b), hence similar (and limited) computational performance due to fully connected layers.*	87.2% mIoU @PASCAL 2012	★ ★

Table 1. (Continued).

Method	Method Summary	Rankings	Eff.
YOLACT (Bolya et al. 2019)	*Object Detector YOLO followed by Class Probability and Convolutional layers, applied to instance segmentation (see Figure 5.b), with real-time semantic segmentation performance.*	72.3% mAP $_{50}$ @PASCAL SBD 31.2% mAP @COCO	★ ★ ★ ★
ESE-Seg (Xu et al. 2019a)	*ESE-Seg is an object detection-based approach that uses explicit shape encoding by explicitly decoding the multiple object shapes with tensor operations in real-time.*	69.3% mAP $_{50}$ @PASCAL SBD 21.6% mAP @COCO	★ ★ ★ ★
SOLO (Wang et al. 2019a)	*The central idea of SOLO framework is to reformulate the instance segmentation as two simultaneous problems: category prediction and instance mask generation, using a single convolutional backbone. The model can run in real-time with proper parameter tuning.*	37.8% mAP @COCO	★ ★ ★
EfficientNet-L2 + NASFPN + Noisy Student (Zoph et al. 2020)	*The study aims at understanding the effect of pre- and self training and apply this to semantic segmentation problem. For their experiment, they utilize a neural architecture search (NAS) strategy (Ghiasi, Lin, and Le 2019) with EfficientNet-L2 (Xie et al. 2020b) as the backbone architecture. The model is the leader of PASCAL VOC 2012 challenge by the time this manuscript was written.*	90.5% mIoU @PASCAL 2012	★ ★ ★
DCNAS (Zhang et al. 2020b)	*Neural Architecture Search applied to MobileNetV3 (Howard et al. 2019), a densely connected search space for semantic segmentation. Although computational performance is not explicitly indicated, the resulting architecture possibly provides U-Net like computational efficiency for model inference.*	86.9% mIoU @PASCAL 2012 (taken from the paper, not listed in the leaderboard) 83.6% mIoU @CitySca. (w. course)	★ ★ ★
SOLOv2 (Wang et al. 2020)	*Updated, real-time version of SOLO (Wang et al. 2019a), empowered by an efficient and holistic instance mask representation scheme, which dynamically segments each instance in the image, without resorting to bounding box detection.*	37.1% mAP @COCO	★ ★ ★ ★
Deep Snake (Peng et al. 2020)	*Deep Snake is a fully convolutional architecture with a contour-based approach for real-time instance segmentation.*	62.1% mAP $_{50}$ @PASCAL SBD 30.3% mAP @COCO	★ ★ ★ ★
BlendMask (Chen et al. 2020)	*Using both top-down and bottom-up instance segmentation approaches, BlendMask learns attention maps for each instance using a single convolution layer.*	37.1% mAP @COCO	★ ★ ★ ★
SwiftNetRN18-Pyr (Oršić and Šegvić 2021)	*Based on shared pyramidal representation and fusion of heterogeneous features, SwiftNetRN18-Pry fuses hybrid representation within a ladder-style decoder. Provides beyond real-time performance with modest accuracy.*	35.0% mIoU @ADE20K	★ ★ ★ ★
BOXInst (Tian et al. 2021)	*Achieves mask-level instance segmentation with only bounding-box annotations for training. Core idea is to redesign the loss of learning masks in instance segmentation*	61.4% mAP $_{50}$ @PASCAL SBD 31.6% mAP @COCO	★ ★ ★

Pre-Deep Learning Approaches

The differentiating factor between conventional image segmentation and semantic segmentation is the utilization of semantic features in the process. Conventional methods for image segmentation such as thresholding, clustering, and region growing, etc., (*please see* (Zaitoun and Aqel 2015) *for a survey on conventional image segmentation techniques*) utilize handcrafted low-level features (i.e., edges and blobs) to locate object boundaries in images. Thus, in situations where the semantic information of an image is necessary for pixel-wise segmentation, such as in similar objects occluding each other, these methods usually return a poor performance.

Regarding semantic segmentation efforts prior to DCNNs becoming popular, a wide variety of approaches (Fröhlich, Rodner, and Denzler 2013; He and Zemel 2008; Krähenbühl and Koltun 2011; Ladický et al. 2009; Lempitsky, Vedaldi, and Zisserman 2011; Mičušĺík and Košecká 2009; Montillo et al. 2011; Rav et al. 2016; Shotton, Johnson, and Cipolla 2008; Ulusoy and Bishop 2005; Vezhnevets, Ferrari, and Buhmann 2011; Xiao and Quan 2009; Yao, Fidler, and Urtasun 2012) utilized graphical models, such as Markov Random Fields (MRF), Conditional Random Fields (CRF) or forest-based (or sometimes referred to as 'holistic') methods, in order to find scene labels at the pixel level. The main idea was to find an inference by observing the dependencies between neighboring pixels. In other words, these methods modeled the semantics of the image as a kind of prior information among adjacent pixels. Thanks to deep learning, today we know that image semantics require abstract exploitation of large-scale data. Initially, graph-based approaches were thought to have this potential. The so-called 'super-pixelisation,' which is usually the term applied in these studies, was a process of modeling abstract regions. However, a practical and feasible implementation for large-scale data processing was never achieved for these methods, while it was accomplished for DCNNs, first by (Krizhevsky et al. 2012) and then in many other studies.

Another group of studies, sometimes referred to as the 'Layered models' (Arbeláez et al. 2012; Ladický et al. 2010; Yang et al. 2012), used a composition of pre-trained and separate object detectors so as to extract the semantic information from the image. Because the individual object detectors failed to classify regions properly, or because the methods were limited by the finite number of object classes provided by the 'hand-selected' bank of detectors in general, their performance was seen as relatively low compared to today's state-of-the-art methods.

Although the aforementioned methods of the pre-deep learning era are no longer preferred as segmentation methods, some of the graphical models, especially CRFs, are currently being utilized by the state-of-the-art methods

as post-processing (refinement) layers, with the purpose of improving the semantic segmentation performance, the details of which are discussed in the following section.

Refinement Methods

Deep neural networks are powerful in extracting abstract local features. However, they lack the capability to utilize global context information, and accordingly cannot model interactions between adjacent pixel predictions (Teichmann and Cipolla 2018). On the other hand, the popular segmentation methods of the pre-deep learning era, the graphical models, are highly suited to this sort of task. That is why they are currently being used as a refinement layer on many DCNN-based semantic segmentation architectures.

As also mentioned in the previous section, the idea behind using graphical models for segmentation is finding an inference by observing the low-level relations between neighboring pixels. In Figure 2, the effect of using a graphical model-based refinement on segmentation results can be seen. The classifier (see Figure 2.b) cannot correctly segment pixels where different class labels are adjacent. In this example, a CRF-based refinement (Krähenbühl and Koltun 2011) is applied to improve the pixel-wise segmentation results. CRF-based methods are widely used for the refinement of deep semantic segmentation methods, although some alternative graphical model-based refinement methods also exist in the literature (Liu et al. 2015; Zuo and Drummond 2017).

CRFs (Lafferty, McCallum, and Pereira 2001) are a type of discriminative undirected probabilistic graphical model. They are used to encode known relationships between observations and to construct consistent interpretations. Their usage as a refinement layer comes from the fact that, unlike a discrete classifier, which does not consider the similarity of adjacent pixels, a CRF can utilize this information. The main advantage of CRFs over other graphical models (such as Hidden Markov Models) is their conditional nature and their ability to avoid the problem of label bias (Lafferty, McCallum, and Pereira 2001). Even though a considerable number of methods (see Table 1) utilize CRFs for refinement, these models started to lose popularity in relatively recent approaches because they are notoriously slow and very difficult to optimize (Teichmann and Cipolla 2018).
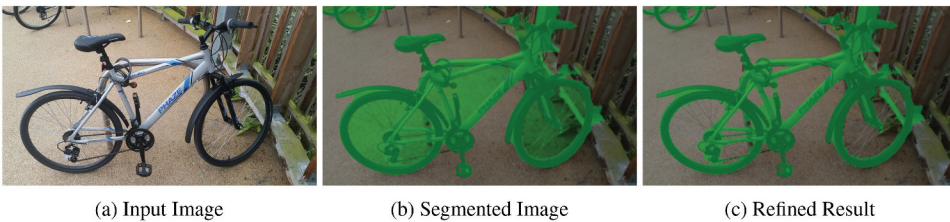

(a) Input Image   (b) Segmented Image   (c) Refined Result

Figure 2. Effect of using graphical model-based refinement on segmentation results.

Early Deep Learning Approaches

Before FCNs first appeared in 2014,[6] the initial few years of deep convolutional networks saw a growing interest in the idea of utilizing the newly discovered deep features for semantic segmentation (Ciresan et al. 2012; Farabet et al. 2013; Ganin and Lempitsky 2014; Hariharan et al. 2014; Ning et al. 2005; Pinheiro and Collobert 2014). The very first approaches, which were published prior to the proposal of a Rectified Linear Unit (ReLU) layer (Krizhevsky et al. 2012), used activation functions such as *tanh* (Ning et al. 2005) (or similar continuous functions), which can be computationally difficult to differentiate. Thus, training such systems were not considered to be computation-friendly, or even feasible for large-scale data.

However, the first mature approaches were just simple attempts to convert classification networks such Alex-Net and VGG to segmentation networks by fine-tuning the fully connected layers (Ciresan et al. 2012; Ganin and Lempitsky 2014; Ning et al. 2005). They suffered from the overfitting and time-consuming nature of their fully connected layers in the training phase. Moreover, the CNNs used were not sufficiently deep so as to create abstract features, which would relate to the semantics of the image.

There were a few early deep learning studies in which the researchers declined to use fully connected layers for their decision-making. However, they utilized different structures such as a recurrent architecture (Pinheiro and Collobert 2014) or using labelling from a family of separately computed segmentations (Farabet et al. 2013). By proposing alternative solutions to fully connected layers, these early studies showed the first traces of the necessity for a structure like the FCN, and unsurprisingly they were succeeded by (Shelhamer, Long, and Darrell 2017).

Since their segmentation results were deemed to be unsatisfactory, these studies generally utilized a refinement process, either as a post-processing layer(Ciresan et al. 2012; Ganin and Lempitsky 2014; Hariharan et al. 2014; Ning et al. 2005) or as an alternative architecture to fully connected decision layers (Farabet et al. 2013; Pinheiro and Collobert 2014). Refinement methods varied, such as Markov random fields (Ning et al. 2005), nearest neighbor-based approach (Ganin and Lempitsky 2014), the use of a calibration layer (Ciresan et al. 2012), using super-pixels (Farabet et al. 2013; Hariharan et al. 2014), or a recurrent network of plain CNNs (Pinheiro and Collobert 2014). Refinement layers, as discussed in the previous section, are still being utilized by post-FCN methods, with the purpose of increasing the pixel-wise labelling performance around regions where class intersections occur.

Fully Convolutional Networks for Semantic Segmentation

In (Shelhamer, Long, and Darrell 2017), the idea of dismantling fully connected layers from deep CNNs (DCNN) was proposed, and to imply this idea, the proposed architecture was named as 'Fully Convolutional Networks' (see Figure 3). The main objective was to create semantic segmentation networks by adapting classification networks such as AlexNet (Krizhevsky et al. 2012), VGG (Simonyan and Zisserman 2015), and GoogLeNet (Szegedy et al. 2015) into fully convolutional networks, and then transferring their learnt representations by fine-tuning. The most widely used architectures obtained from the study (Shelhamer, Long, and Darrell 2017) are known as 'FCN-32s,' 'FCN16s,' and 'FCN8s,' which are all transfer-learnt using the VGG architecture (Simonyan and Zisserman 2015).

FCN architecture was considered revolutionary in many aspects. First of all, since FCNs did not include fully connected layers, inference per image was seen to be considerably faster. This was mainly because convolutional layers when compared to fully connected layers, had a marginal number of weights. Second, and maybe more significant, the structure allowed segmentation maps to be generated for images of any resolution. In order to achieve this, FCNs used deconvolutional layers that can upsample coarse deep convolutional layer outputs to dense pixels of any desired resolution. Finally, and most importantly, they proposed the skip architecture for DCNNs.

Skip architectures (or connections) provide links between nonadjacent layers in DCNNs. Simply by summing or concatenating outputs of unconnected layers, these connections enable information to flow, which would otherwise be lost because of an architectural choice such as max-pooling layers or dropouts. The most common practice is to use skip connections preceding a max-pooling layer, which downsamples layer output by choosing the maximum value in a specific region. Pooling layers helps the architecture create feature hierarchies, but also causes loss of localized information, which could
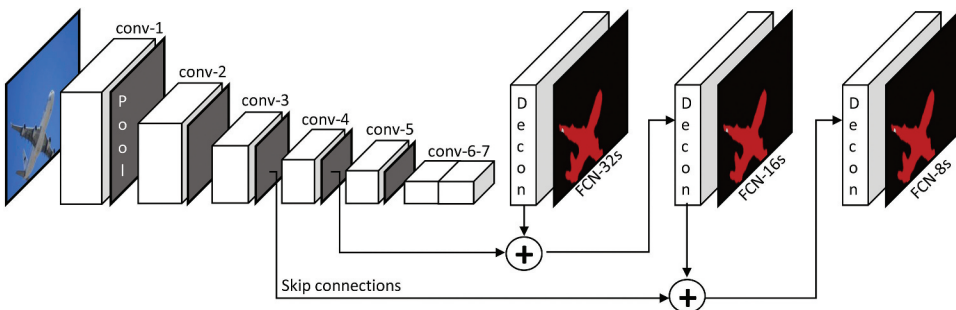

Figure 3. Fully convolutional networks (FCNs) are trained end-to-end and are designed to make dense predictions for per-pixel tasks like semantic segmentation. FCNs consist of no fully connected layers.

be valuable for semantic segmentation, especially at object borders. Skip connections preserve and forward this information to deeper layers by way of bypassing the pooling layers. Actually, the usage of skip connections in (Shelhamer, Long, and Darrell 2017) was perceived as being considerably primitive. The 'FCN-8s' and 'FCN-16s' networks included these skip connections at different layers. Denser skip connections for the same architecture, namely 'FCN-4s' and 'FCN-2s,' were also utilized for various applications (Lee et al. 2017; Zhong et al. 2016). This idea eventually evolved into the encoder-decoder structures (Badrinarayanan, Kendall, and Cipolla 2015; Ronneberger, Fischer, and Brox 2015) for semantic segmentation, which are presented in the following section.

Post-FCN Approaches

Almost all subsequent approaches on semantic segmentation followed the idea of FCNs; thus it would not be wrong to state that decision-making with fully connected layers effectively ceased to exist[7] following the appearance of FCNs to the issue of semantic segmentation.

On the other hand, the idea of FCNs also created new opportunities to further improve deep semantic segmentation architectures. Generally speaking, the main drawbacks of FCNs can be summarized as inefficient loss of label localization within the feature hierarchy, inability to process global context knowledge, and the lack of a mechanism for multiscale processing. Thus, most subsequent studies have been principally aimed at solving these issues through the proposal of various architectures or techniques. For the remainder of this paper, we analyze these issues under the title, 'fine-grained localisation.' Consequently, before presenting a list of the post-FCN state-of-the-art methods, we focus on this categorization of techniques and examine different approaches that aim at solving these main issues. In the following, we also discuss scale invariance in the semantic segmentation context and finish with object detection-based approaches, which are a new breed of solution that aim at resolving the semantic segmentation problem simultaneously with detecting object instances.

Techniques for Fine-grained Localisation

Semantic segmentation is, by definition, a dense procedure; hence, it requires fine-grained localisation of class labels at the pixel level. For example, in robotic surgery, pixel errors in semantic segmentation can lead to life or death situations. Hierarchical features created by pooling (i.e., max-pooling) layers can partially lose localisation. Moreover, due to their fully convolutional nature, FCNs do not inherently possess the ability to model global context information in an image, which is also very effective in the localisation of class

labels. Thus, these two issues are intertwined in nature, and in the following, we discuss different approaches that aim at overcoming these problems and to provide finer localisation of class labels.

Encoder-Decoder Architecture

The so-called Encoder-Decoder (ED) architectures (also known as the U-nets, referring to the pioneering study of (Ronneberger, Fischer, and Brox 2015)) are comprised of two parts. Encoder gradually reduces the spatial dimension with pooling layers, whilst the decoder gradually recovers the object details and spatial dimension. Each feature map of the decoder part only directly receives the information from the feature map at the same level of the encoder part using skip connections, thus EDs can create abstract hierarchical features with fine localisation (see Figure 4.a). U-Net (Ronneberger, Fischer, and Brox 2015) and Seg-Net (Badrinarayanan, Kendall, and Cipolla 2015) are very well-known examples. In this architecture, the strongly correlated semantic information, which is provided by the adjacent lower-resolution feature map of the encoder part, has to pass through additional intermediate layers in order to reach the same decoder layer. This usually results in a level of information decay. However, U-Net architectures have proven very useful for the segmentation of different applications, such as medical images (Ronneberger, Fischer, and Brox 2015), street view images (Badrinarayanan, Kendall, and Cipolla 2015), satellite images (Ulku et al. 2019), just to name a few. Although earlier

(a) Encoder-Decoder Architecture.

(b) Spatial-Pyramid Pooling Layer

(c) Regular vs. Dilated Convolutions

(d) DeepLabv3+ Architecture
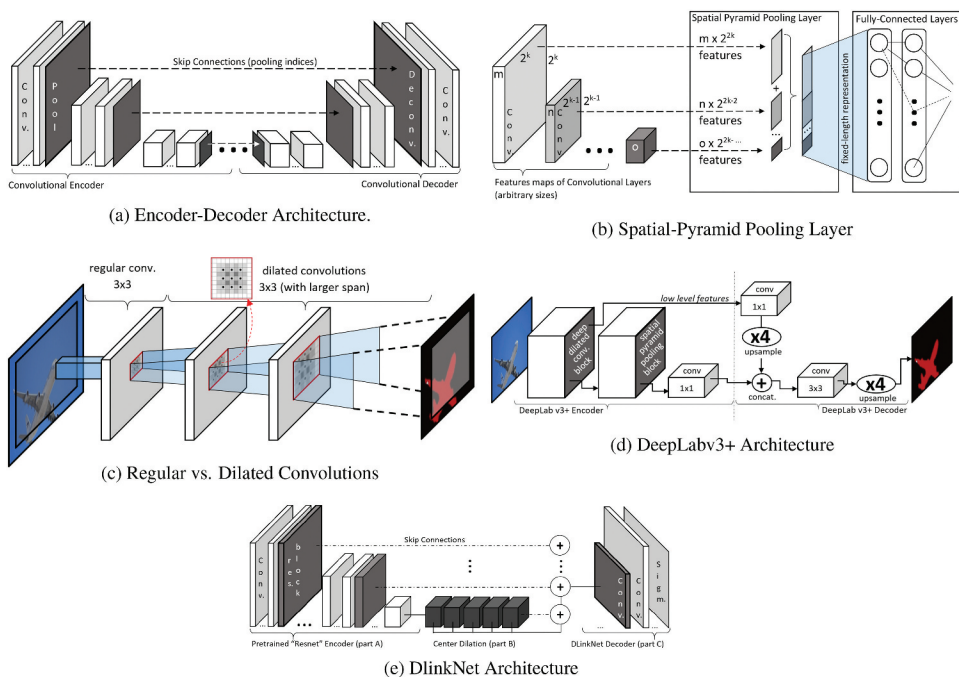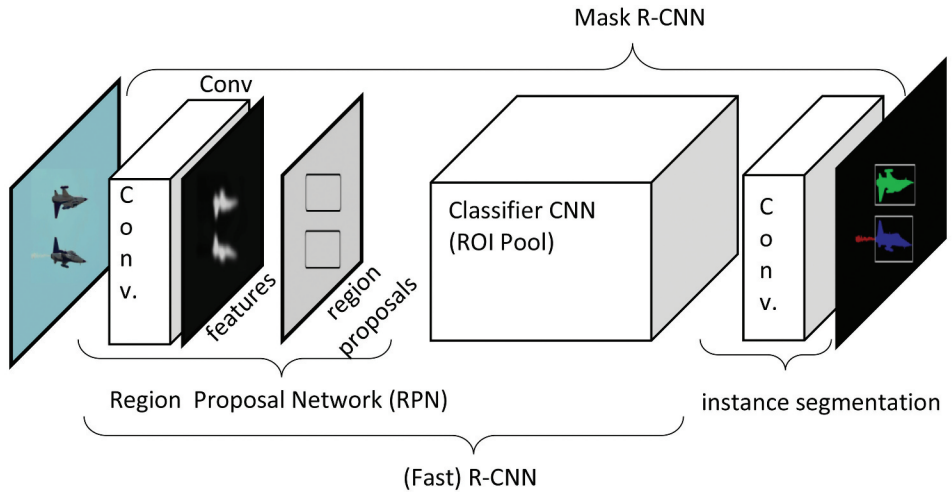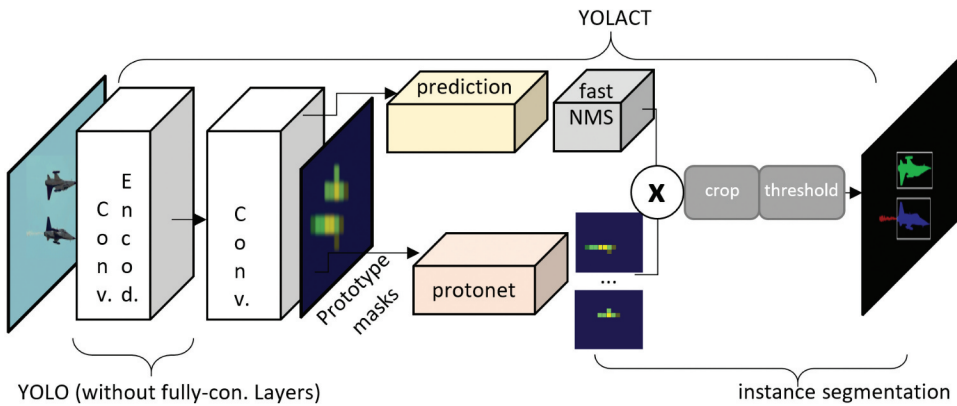
(e) DlinkNet Architecture

Figure 4. Different architectures for fine-grained pixel-wise label localisation.

(a) Regions with CNN features-based (Mask-RCNN) architecture

(b) Fully Convolutional Object Detector-based (YOLO)-based architecture

Figure 5. Different architectures for object detection-based semantic segmentation methods.

ED architectures were designed for object segmentation tasks only, there are also modified versions such as "TernausNetV2" (Iglovikov et al. 2018), that provide instance segmentation capability with minor architectural changes.

Spatial Pyramid Pooling

The idea of constructing a fixed-sized spatial pyramid was first proposed by (Lazebnik, Schmid, and Ponce 2006), in order to prevent a Bag-of-Words system losing spatial relations among features. Later, the approach was adopted to CNNs by (He et al. 2015), in that, regardless of the input size, a spatial pyramid representation of deep features could be created in a Spatial Pyramid Pooling Network (SPP-Net). The most important contribution of the SPP-Net was that it allowed inputs of different sizes to be fed into CNNs.

Images of different sizes fed into convolutional layers inevitably create different-sized feature maps. However, if a pooling layer, just prior to a decision layer, has stride values proportional to the input size, the feature map created by that layer would be fixed (see Figure 4.b). By (Li et al. 2018), a modified version, namely Pyramid Attention Network (PAN) was additionally proposed. The idea of PAN was combining an SPP layer with global pooling to learn a better feature representation.

There is a common misconception that SPP-Net structure carries an inherent scale-invariance property, which is incorrect. SPP-Net allows the efficient training of images at different scales (or resolutions) by allowing different input sizes to a CNN. However, the trained CNN with SPP is scale-invariant if, and only if, the training set includes images with different scales. This fact is also true for a CNN without SPP layers.

However, similar to the original idea proposed in (Lazebnik, Schmid, and Ponce 2006), the SPP layer in a CNN constructs relations among the features of different hierarchies. Thus, it is quite similar to skip connections in ED structures, which also allow information flow between feature hierarchies. The most common utilization of an SPP layer for semantic segmentation is proposed in (He et al. 2015), such that the SPP layer is appended to the last convolutional layer and fed to the pixel-wise classifier.

Feature Concatenation

This idea is based on fusing features extracted from different sources. For example, in (Pinheiro, Collobert, and Dollar 2015) the so-called 'DeepMask' network utilizes skip connections in a feed-forward manner, so that an architecture partially similar to both SPP layer and ED is obtained. The same group extends this idea with a top-down refinement approach of the feed-forward module and propose the so-called 'SharpMask' network (Pinheiro et al. 2016), which has proven to be more efficient and accurate in segmentation performance. Another approach from this category is the so-called 'ParseNet' (Liu et al. 2015), which fuses CNN features with external global features from previous layers in order to provide context knowledge. Another approach by (Wang et al. 2020) is to fuse the "stage features" (i.e. deep encoder activations) with "refinement path features" (an idea similar to skip connections), using a convolutional (Feature Adaptive Fusion FAF) block. Although a novel idea in principle, feature fusion approaches (including SPP) create hybrid structures; therefore, they are relatively difficult to train.

Dilated Convolution

The idea of dilated (atrous) convolutions is actually quite simple: with contiguous convolutional filters, an effective receptive field of units can only grow linearly with layers; whereas with dilated convolution, which has gaps in the filter (see Figure 4.c), the effective receptive field would grow much more

quickly (Chen et al. 2018a). Thus, with no pooling or subsampling, a rectangular prism of convolutional layers is created. Dilated convolution is a very effective and powerful method for the detailed preservation of feature map resolutions. The negative aspect of the technique, compared to other techniques, concerns its higher demand for GPU storage and computation, since the feature map resolutions do not shrink within the feature hierarchy (He et al. 2016).

Conditional Random Fields

As also discussed in Section 3.1.1, CNNs naturally lack mechanisms to specifically 'focus' on regions where class intersections occur. Around these regions, graphical models are used to find inference by observing low-level relations between neighboring feature maps of CNN layers. Consequently, graphical models, mainly CRFs, are utilized as refinement layers in deep semantic segmentation architectures. As in (Rother, Kolmogorov, and Blake 2004), CRFs connect low-level interactions with output from multiclass interactions, and in this way, global context knowledge is constructed.

As a refinement layer, various methods exist that employ CRFs to DCNNs, such as the Convolutional CRFs (Teichmann and Cipolla 2018), the Dense CRF (Krähenbühl and Koltun 2011), and CRN-as-RNN (Zheng et al. 2015). In general, CRFs help build context knowledge and thus a finer level of localisation in class labels.

Recurrent Approaches

The ability of Recurrent Neural Networks (RNNs) to handle sequential information can help improve segmentation accuracy. For example, (Pfeuffer, Schulz, and Dietmayer 2019) used Conv-LSTM layers to improve their semantic segmentation results in image sequences. However, there are also methods that use recurrent structures on still images. For example, the Graph LSTM network (Liang et al. 2016) is a generalization of LSTM from sequential data or multidimensional data to general graph-structured data for semantic segmentation on 2D still images. Graph-RNN Shuai2016 is another example of a similar approach in which an LSTM-based network is used to fuse a deep encoder output with the original image in order to obtain a finer pixel-level segmentation. Likewise, in (Lin et al. 2018), the researchers utilized LSTM-chains in order to intertwine multiple scales, resulting in pixel-wise segmentation improvements. There are also hybrid approaches where CNNs and RNNs are fused. A good example of this is the so-called ReSeg model (Visin et al. 2016), in which the input image is fed to a VGG-like CNN encoder, and is then processed afterwards by recurrent layers (namely the ReNet architecture) in order to better localize the pixel labels. Another similar approach is the DAG-RNN (Shuai et al. 2016), which utilize a DAG-structured CNN+RNN network, and models long-range semantic dependencies among image units. To the best

of our knowledge, no purely recurrent structures for semantic segmentation exist, mainly because semantic segmentation requires a preliminary CNN-based feature encoding scheme.

There is currently an increasing trend in one specific type of RNN, namely 'recurrent attention modules.' In these modules, attention (Vaswani et al. 2017) is technically fused in the RNN, providing a focus on certain regions of the input when predicting a certain part of the output sequence. Consequently, they are also being utilized in semantic segmentation (Li et al. 2019b; Zhao et al. 2018b; Oktay et al. 2018).

Scale-Invariance

Scale Invariance is, by definition, the ability of a method to process a given input, independent of the relative scale (i.e. the scale of an object to its scene) or image resolution. Although it is extremely crucial for certain applications, this ability is usually overlooked or is confused with a method's ability to include multiscale information. A method may use multiscale information to improve its pixel-wise segmentation ability, but can still be dependent on scale or resolution. That is why we find it necessary to discuss this issue under a different title and to provide information on the techniques that provide scale and/or resolution invariance.

In computer vision, any method can become scale invariant if trained with multiple scales of the training set. Some semantic segmentation methods, such as (Eigen and Fergus 2014; Farabet et al. 2013; Lin et al. 2016a; Pinheiro and Collobert 2014; Yu and Koltun 2015) utilize this strategy. However, these methods do not possess an inherent scale-invariance property, which is usually obtained by normalization with a global scale factor (such as in SIFT by (Lowe 2004)). This approach is not usually preferred in the literature on semantic segmentation. The image sets that exist in semantic segmentation literature are extremely large in size. Thus, the methods are trained to memorize that training set, because in principle, overfitting a large-scale training set is actually tantamount to solving the entire problem space.

Object Detection-based Methods

There has been a recent growing trend in computer vision, which aims at specifically resolving the problem of object detection, that is, establishing a bounding box around all objects within an image. Given that the image may or may not contain any number of objects, the architectures utilized to tackle such a problem differ to the existing fully-connected/convolutional classification or segmentation models.

The pioneering study that represents this idea is the renowned 'Regions with CNN features' (RCNN) network (Girshick et al. 2013). Standard CNNs with fully convolutional and fully connected layers lack the ability to provide varying length output, which is a major flaw for an object detection algorithm that aims to detect an unknown number of objects within an image. The simplest way to resolve this problem is to take different regions of interest from the image, and then to employ a CNN in order to detect objects within each region separately. This region selection architecture is called the 'Region Proposal Network' (RPN) and is the fundamental structure used to construct the RCNN network (see Figure 5.a). Improved versions of RCNN, namely 'Fast-RCNN' (Girshick et al. 2013) and 'Faster-RCNN' (Ren et al. 2015) were subsequently also proposed by the same research group. Because these networks allow for the separate detection of all objects within the image, the idea was easily implemented for instance segmentation, as the 'Mask-RCNN' (He et al. 2017).

The basic structure of RCNNs included the RPN, which is the combination of CNN layers and a fully connected structure in order to decide the object categories and bounding box positions. As discussed within the previous sections of this paper, due to their cumbersome structure, fully connected layers were largely abandoned with FCNs. RCNNs shared a similar fate when the 'You-Only-Look-Once' (YOLO) by (Redmon et al. 2016) and 'Single Shot Detector' (SSD) by (Liu et al. 2016) were proposed. YOLO utilizes a single convolutional network that predicts the bounding boxes and the class probabilities for these boxes. It consists of no fully connected layers, and consequently provides real-time performance. SSD proposed a similar idea, in which bounding boxes were predicted after multiple convolutional layers. Since each convolutional layer operates at a different scale, the architecture is able to detect objects of various scales. Whilst slower than YOLO, it is still considered to be faster then RCNNs. This new breed of object detection techniques was immediately applied to semantic segmentation. Similar to MaskRCNN, 'Mask-YOLO' (Sun 2019) and 'YOLACT' (Bolya et al. 2019) architectures were implementations of these object detectors to the problem of instance segmentation (see Figure 5b). Similar to YOLACT, some other methods also achieve fast, real-time instance segmentation such as: ESE-Seg (Xu et al. **2019a**), SOLO (Wang et al. 2019a), SOLOv2 (Wang et al. **2020a**), DeepSnake (Peng et al. 2020), and CenterPoly (Perreault et al. 2021).

Locating objects within an image prior to segmenting them at the pixel level is both intuitive and natural, due to the fact that it is effectively how the human brain supposedly accomplishes this task (Rosenholtz 2016). In addition to these two-stage (detection+segmentation) methods, there are some recent studies that aim at utilizing the segmentation task to be incorporated into one-stage bounding-box detectors and result in a simple yet efficient instance segmentation framework (Xu et al. **2019a**; R. Zhang et al. **2020a**; Lee and

Park 2020; Xie et al. 2020a). However, the latest trend is to use global-area-based methods by generating intermediate FCN feature maps and then assembling these basis features to obtain final masks (Chen et al. 2020; Ke, Tai, and Tang 2021; Kim et al. 2021).

In recent years, a trend of alleviating the demand for pixel-wise labels is realized mainly by employing bounding boxes, and by expanding from semantic segmentation to instance segmentation applications. In both semantic segmentation and instance segmentation methods, the category of each pixel is recognized, and the only difference is that instance segmentation also differentiates object occurrences of the same category. Therefore, weakly-supervisedinstance segmentation (WSIS) methods are also utilized for instance segmentation. The supervision of WSIS methods can use different annotation types for training, which are usually in the form of either bounding boxes (Khoreva et al. 2017; Hsu et al. 2019; Arun, Jawahar, and Kumar 2020; Tian et al. 2021; Lee et al. 2021; Cheng et al. **2021b**) or image-level labels (Liu et al. 2020; Shen et al. **2021b**; Zhou et al. 2016; Shen et al. **2021a**). Hence, employing object detection-based methods for semantic segmentation is an area significantly prone to further development in near future by the time this manuscript is prepared.

Evolution of Methods

In Table 1, we present several semantic segmentation methods, each with a brief summary, explaining the fundamental idea that represents the proposed solutions, their position in available leaderboards, and a categorical level of the method's computational efficiency. The intention is for readers to gain a better evolutionary understanding of the methods and architectures in this field, and a clearer conception of how the field may subsequently progress in the future. Regarding the brief summaries of the listed methods, please refer to the categorizations provided earlier in this section.

Table 1 includes 34 methods spanning an eight-year period, starting with early deep learning approaches through to the most recent state-of-the-art techniques. Most of the listed studies have been quite successful and have significantly high rankings in the previously mentioned leaderboards. Whilst there are many other methods, we believe this list to be a clear depiction of the advances in deep learning-based semantic segmentation approaches. In Figure 6, a sample image from the PASCAL VOC validation set, its semantic segmentation ground truth and results obtained from some of the listed studies are depicted. Figure 6 clearly shows the gradually growing success of different methods starting with the pioneering FCN architectures to more advanced architectures such as DeepLab (Chen et al. **2014a**, Chen et al. 2018a) or CRF-as-RNN (Zheng et al. 2015).
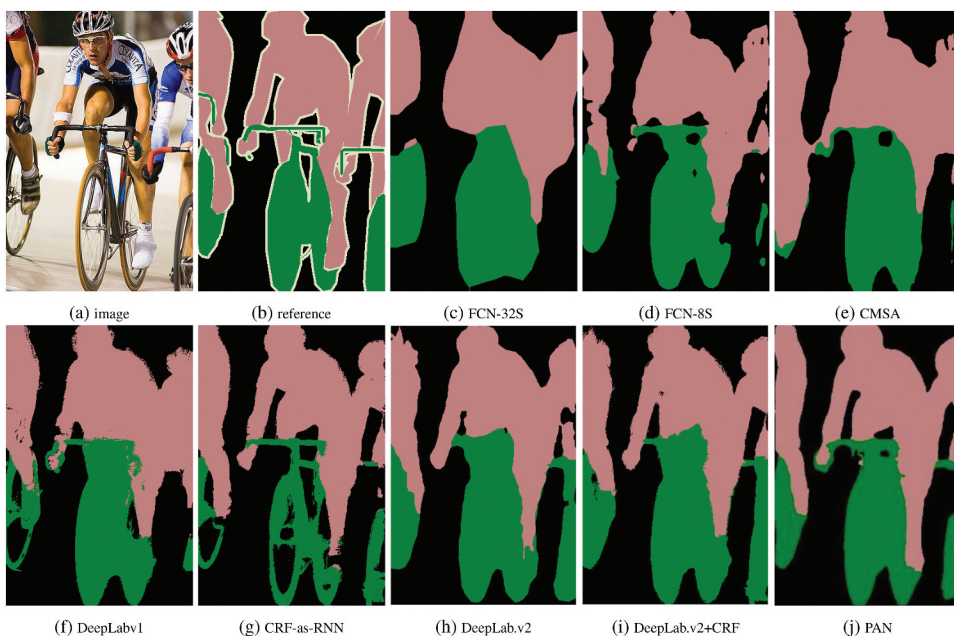
Figure 6. (a) A sample image from the PASCAL VOC validation set, (b) its semantic segmentation ground truth, and results obtained from different studies are depicted: c) FCN-32S (Shelhamer, Long, and Darrell 2017), d) FCN-8S (Shelhamer, Long, and Darrell 2017), e) CMSA (Eigen and Fergus 2014), f) DeepLab-v1 (Chen et al. **2014a**), g) CRF-as-RNN (Zheng et al. 2015), h) DeepLab-v2 (Chen et al. 2018a), i) DeepLab-v2 with CRF refinement (Chen et al. 2018a), (j) PAN (Li et al. 2018).

Judging by the picture it portrays, the deep evolution of the literature clearly reveals a number of important implications. First, graphical model-based refinement modules are being abandoned due to their slow nature. A good example of this trend would be the evolution of DeepLab from (Chen et al. **2014a**) to (Chen et al. 2018b) (see Table 1). Notably, no significant study published in 2019 and 2020 employed a CRF-based or similar module to refine their segmentation results. Second, most studies published in the past two years show no significant leap in performance rates. For this reason, researchers have tended to focus on experimental solutions such as object detection-based or Neural Architecture Search (NAS)-based approaches. Some of these very recent group of studies (Zhang et al. 2020b; Zoph et al. 2020) focus on (NAS)-based techniques, instead of hand-crafted architectures. EfficientNet-NAS (Zoph et al. **2020**) belongs to this category and is the leading study in PASCAL VOC 2012 semantic segmentation challenge at the time the paper was prepared. We believe that the field will witness an increasing interest in NAS-based methods in the near future. In general, considering all studies of the post-FCN era, the main challenge of the field still remains to be *efficiently* integrating (i.e. in real-time) global context to localisation information, which still does not appear to have an off-the-shelf solution, although there are some promising techniques, such as YOLACT (Bolya et al. 2019).

In Table 1, the right-most column represents a categorical level of computational efficiency. We use a four-level categorization (one star to four stars) to indicate the computational efficiency of each listed method. For any assigned level of the computational efficiency of a method, we explain our reasoning in the table with solid arguments. For example, one of the four-star methods in Table 1 is "YOLACT" by (Bolya et al. 2019), which claims to provide real-time performance (i.e. > 30fps) on both PASCAL VOC 2012 and COCO image sets.

Future Scope and Potential Research Directions

Although tremendous successes have been achieved so far in the semantic segmentation field, there are still many open challenges in this field due to hard requirements time-consuming pixel-level annotations, lack of generalization ability to new domains and classes, and need for real-time performance with higher segmentation accuracies. In this section, we categorize possible future directions under different titles by providing examples of recent studies that represent that direction.

Weakly-Supervised Semantic Segmentation (WSSS)

Over the last few years, there has been an increasing research effort directed toward the approaches that are alternative to pixel-level annotations such as; unsupervised, semi-supervised (He, Yang, and Qi 2021) and weakly-supervised methods. Recent studies show that, WSSS methods usually perform better than the other schemes (Chan, Hosseini, and Plataniotis 2021) where annotations are in the form of image-level labels (Kolesnikov and Lampert 2016; Pathak, Krahenbuhl, and Darrell 2015; Pinheiro and Collobert 2015; Wang et al. **2020b**; Ahn and Kwak 2018; Li et al. **2021b**; Chang et al. 2020; Xu et al. 2021; Yao et al. 2021; Jiang et al. 2021), video-level labels (Zhong et al. 2016b), scribbles (Lin et al. 2016a), points (Bearman et al. 2016), and bounding boxes (Dai, He, and Sun 2015; Khoreva et al. 2017; Xu, Schwing, and Urtasun 2015). In case of image-level labels, class activation maps (CAMs) (Zhou et al. 2016) are used to localize the small discriminative regions, which are not suitable particularly for the large-scale objects, but can be utilized as initial seeds (pseudo-masks) (Araslanov and Roth 2020; Fan et al. 2020; Kweon et al. 2021; Sun et al. 2021).

Zero-/Few-Shot Learning

Motivated by humans' ability to recognize new concepts in a scene by using only a few visual samples, zero-shot and/or few-shot learning methods have been introduced. Few-shot semantic segmentation (FS3) methods (Wang et al.

2019a; Xie et al. 2021) has been proposed to recognize objects from unseen classes by utilizing few annotated examples; however, these methods are limited to handling a single unseen class only. Zero-shot semantic segmentation (ZS3) methods have been developed recently to generate visual features by exploiting word embedding vectors in the case of zero training samples (Bucher et al. 2019; Lu et al. 2021; Pastore et al. 2021; Xian et al. 2019). However, the major drawback of ZS3 methods is their insufficient prediction ability to distinguish between the seen and the unseen classes even if both are included in a scene. This disadvantage is usually overcome by generalized ZS3 (GZS3), which recognizes both seen and unseen classes simultaneously. GZS3 studies mainly rely on generative-based methods. Feature extractor training is realized without considering semantic features in GZS3 adopted with generative approaches so that the bias is introduced toward the seen classes. Therefore, GZS3 methods result in performance reduction on unseen classes (Pastore et al. 2021). Much of the recent work on ZS3 has involved such as: exploiting joint embedding space to alleviate the seen bias problem (Baek, Oh, and Ham 2021), analyzing different domain performances chan2021comprehensive, and incorporating spatial information (Cheng et al. **2021a**).

Domain Adaptation

Recent studies also rely on the use of synthetic large-scale image sets such as GTA5 (Richter et al. 2016) and SYNTHIA (Ros et al. 2016b) because of their capability to cope with laborious pixel-level annotations. Although these rich-labeled synthetic images have the advantage of reducing the labeling cost, they also bring about domain shift while training with unlabeled real images. Therefore, applying domain adaptation for aligning the synthetic and the real image sets is of much importance (Zhao et al. 2019; Kang et al. 2020; Wu et al. 2021; Wang et al.2021; Shin et al. 2021; Fleuret et al. 2021). Unsupervised domain adaptation (UDA) methods are widely employed in semantic segmentation (Cheng et al. **2021c**; Liu, Zhang, and Wang 2021; Hong et al. 2018; Vu et al. 2019; Pan et al. 2020; Wang et al. 2021; Saporta et al. 2021; Zheng and Yang 2021).

Real-Time Processing

Adopting compact and shallow model architectures (Zhao et al. **2018a**; Orsic et al. 2019; Yu et al. **2018a**; Li et al. 2019a; Fan et al. 2021) and restricting the input to be low-resolution (Marin et al. 2019) are brand new innovations proposed very recently to overcome the computational burden of large-scale semantic segmentation. To choose a real-time semantic segmentation strategy, all aspects of an application should be considered, as all of these strategies somehow correlate with decreasing the model's discriminative ability and

losing information of object boundaries or small objects to some extent. Some other strategies have also been proposed for the retrieval of rich contextual information in real-time applications including attention mechanisms (Ding et al. 2021; Hu et al. 2020), depth-wise separable convolutions (Chollet 2017; Howard et al. 2019), pyramid fusion (Oršić and Šegvić 2021; Rosas-Arias et al. 2021), grouped convolutions (Zhang et al. **2018b**; Huang et al. 2018) neural architecture search (Zoph et al. 2018) and, pipeline parallelism (Chew, Ji, and Zhang 2022).

Contextual Information

Contextual information aggregation with the purpose of augmenting pixel representations in semantic segmentation architectures is another promising research direction in recent years. In this aspect, mining contextual information (Jin et al. 2021), exploring context information on spatial and channel dimensions (Li et al. **2021c**), focusing on object-based contextual representations (Yuan, Chen, and Wang 2020) and capturing the global contextual information for fine-resolution remote sensing imagery (Li et al. **2021a**) are some of the recent studies. Alternative methods of reducing dense pixel-level annotations in semantic segmentation have been described which are based on using pixel-wise contrastive loss (Chaitanya et al. 2020; Zhang et al. 2021; Zhao et al. 2021).

Conclusions

In this survey, we aimed at reviewing the current developments in the literature regarding deep learning-based 2D image semantic segmentation. We commenced with an analysis of the public image sets and leaderboards for 2D semantic segmentation and then continued by providing an overview of the techniques for performance evaluation. Following this introduction, our focus shifted to the 10-year evolution seen in this field under three chronological titles, namely the pre- and early-deep learning era, the fully convolutional era, and the post-FCN era. After a technical analysis on the approaches of each period, we presented a table of methods spanning all three eras, with a brief summary of each technique that explicates their contribution to the field.

In our review, we paid particular attention to the key technical challenges of the 2D semantic segmentation problem, the deep learning-based solutions that were proposed, and how these solutions evolved as they shaped the advancements in the field. To this end, we observed that the fine-grained localisation of pixel labels is clearly the definitive challenge to the overall problem. Although the title may imply a more 'local' interest, the research published in this field evidently shows that it is the global context that determines the actual performance of a method. Thus, it is eminently conceivable why the literature is rich

with approaches that attempt to bridge local information with a more global context, such as graphical models, context aggregating networks, recurrent approaches, and attention-based modules. It is also clear that efforts to fulfill this local-global semantics gap at the pixel level will continue for the foreseeable future.

Another important revelation from this review has been the profound effect seen from public challenges to the field. Academic and industrial groups alike are in a constant struggle to top these public leaderboards, which has an obvious effect of accelerating development in this field. Therefore, it would be prudent to promote or even contribute to creating similar public image sets and challenges affiliated to more specific subjects of the semantic segmentation problem, such as 2D medical images.

Considering the rapid and continuing development seen in this field, there is an irrefutable need for an update on the surveys regarding the semantic segmentation problem. However, we believe that the current survey may be considered as a milestone in measuring how much the field has progressed thus far, and where the future directions possibly lie.

Notes

1. We consider MRI and CT essentially as 3D volume data. Although individual MRI/CT slices are 2D, when doing semantic segmentation on these types of data, neighbourhood information in all three dimensions are utilised. For this reason, medical applications are excluded from this survey.
2. http://host.robots.ox.ac.uk:8080/leaderboard/main_bootstrap.php
3. http://cocodataset.org
4. http://sceneparsing.csail.mit.edu/
5. https://www.cityscapes-dataset.com/benchmarks/
6. FCN (Shelhamer, Long, and Darrell 2017) was officially published in 2017. However, the same group first shared the idea online as pre-printed literature in 2014 (Long, Shelhamer, and Darrell 2014).
7. Many methods utilise fully connected layers such as RCNN (Girshick 2015), which are discussed in the following sections. However, this and other similar methods that include fully connected layers have mostly been succeeded by fully convolutional versions for the sake of computational efficiency.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Irem Ulku ⓘ http://orcid.org/0000-0003-4998-607X
Erdem Akagündüz ⓘ http://orcid.org/0000-0002-0792-7306

References

Ahmad, T., P. Campr, M. Cadik, and G. Bebis (2017, May). Comparison of semantic segmentation approaches for horizon/sky line detection. *2017 International Joint Conference on Neural Networks (IJCNN)* Anchorage, AK, USA.

Ahn, J., and S. Kwak (2018). Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* Salt Lake City, UT, USA, 4981–90.

Araslanov, N., and S. Roth (2020). Single-stage semantic segmentation from image labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* Seattle, WA, USA, 4253–62.

Arbeláez, P., B. Hariharan, C. Gu, S. Gupta, L. Bourdev, and J. Malik (2012). Semantic segmentation using regions and parts. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* Providence, RI, USA, 3378–85. IEEE.

Arun, A., C. Jawahar, and M. P. Kumar (2020). Weakly supervised instance segmentation by learning annotation consistent instances. In *European Conference on Computer Vision* Glasgow, UK (Virtual/Online), 254–70. Springer.

Badrinarayanan, V., A. Kendall, and R. Cipolla (2015). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *CoRR abs/1511.00561*.

Baek, D., Y. Oh, and B. Ham (2021). Exploiting a joint embedding space for generalized zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* Montreal, Canada, 9536–45.

Bearman, A., O. Russakovsky, V. Ferrari, and L. Fei-Fei (2016). What's the point: Semantic segmentation with point supervision. In *European Conference on Computer Vision* Amsterdam, the Netherlands, 549–65. Springer.

Bolya, D., C. Zhou, F. Xiao, and Y. J. Lee (2019). YOLACT: Real-time instance segmentation. *CoRR abs/1904.02689*.

Brostow, G. J., J. Fauqueur, and R. Cipolla. 2009. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters* 30:88–97. doi:10.1016/j.patrec.2008.04.005.

Bucher, M., T.-H. Vu, M. Cord, and P. Pérez. (2019). Zero-shot semantic segmentation Conference on Neural Information Processing Systems, Vancouver, Canada vol. 32 (Curran Associates, Inc.), 468–479.

Chaitanya, K., E. Erdil, N. Karani, and E. Konukoglu (2020). Contrastive learning of global and local features for medical image segmentation with limited annotations. *arXiv:2006.10511*.

Chan, L., M. S. Hosseini, and K. N. Plataniotis. 2021. A comprehensive analysis of weakly-supervised semantic segmentation in different image domains. *International Journal of Computer Vision* 129 (2):361–84. doi:10.1007/s11263-020-01373-4.

Chang, Y.-T., Q. Wang, W.-C. Hung, R. Piramuthu, Y.-H. Tsai, and M.-H. Yang (2020). Weakly-supervised semantic segmentation via sub-category exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* Seattle, WA, USA, 8991–9000.

Chen, X., R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille (2014b). Detect what you can: Detecting and representing objects using holistic models and body parts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* Columbus, OH, USA.

Chen, L., G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille (2014a). Semantic image segmentation with deep convolutional nets and fully connected crfs. *CoRR abs/1412.7062*.

Chen, L., G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. 2018a. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40(4):834–48. April. doi: 10.1109/TPAMI.2017.2699184.

Chen, L., G. Papandreou, F. Schroff, and H. Adam. 2017. . In Rethinking Atrous Convolution for Semantic Image Segmentation. http://arxiv.org/abs/1706.05587

Chen, H., K. Sun, Z. Tian, C. Shen, Y. Huang, and Y. Yan (2020). Blendmask: Top-down meets bottom-up for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* Seattle, WA, USA, 8573–81.

Chen, L., Y. Zhu, G. Papandreou, F. Schroff, and H. Adam (2018b). Encoder-decoder with atrous separable convolution for semantic image segmentation. *CoRR abs/1802.02611*.

Cheng, J., S. Nandi, P. Natarajan, and W. Abd-Almageed (2021a). Sign: Spatial-information incorporated generative network for generalized zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* Montreal, Canada, 9556–66.

Cheng, B., A. Schwing, and A. Kirillov. 2021b. Per-pixel classification is not all you need for semantic segmentation. In *Advances in Neural Information Processing Systems*, 34.

Cheng, Y., F. Wei, J. Bao, D. Chen, F. Wen, and W. Zhang (2021c). Dual path learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9082–91.

Chew, A. W. Z., A. Ji, and L. Zhang. 2022. Large-scale 3d point-cloud semantic segmentation of urban and rural scenes using data volume decomposition coupled with pipeline parallelism. *Automation in Construction* 133:103995. doi:10.1016/j.autcon.2021.103995.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* Honolulu, HI, USA, 1251–58.

Ciresan, D., A. Giusti, L. M. Gambardella, and J. Schmidhuber. 2012. Deep neural networks segment neuronal membranes in electron microscopy images. In *Advances in neural information processing systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. ed., vol. 25. 2843–51. Curran Associates, Inc.

Cordts, M., M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele (2016). The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition* Las Vegas, NV, USA, 3213–23.

Dai, J., K. He, and J. Sun (2015). Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision* Santiago, Chile, 1635–43.

Defence Science & Technology Laboratory. Dstl Satellite Imagery Feature Detection Dataset. https://www.kaggle.com/c/dstl-satellite-imagery-feature-detection/data. Accessed 3 2 2022.

Ding, X., C. Shen, Z. Che, T. Zeng, and Y. Peng (2021). Scarf: A semantic constrained attention refinement network for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3002–11.

Eigen, D., and R. Fergus (2014). Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. *CoRR abs/1411.4734*.

Everingham, M., L. Gool, C. K. Williams, J. Winn, and A. Zisserman. 2010. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision* 88(2):303–38. June. doi: 10.1007/s11263-009-0275-4.

Fan, M., S. Lai, J. Huang, X. Wei, Z. Chai, J. Luo, and X. Wei (2021). Rethinking bisenet for real-time semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9716–25.

Fan, J., Z. Zhang, C. Song, and T. Tan (2020). Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition*, 4283–92.

Farabet, C., C. Couprie, L. Najman, and Y. LeCun. 2013. Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35 (8):1915–29. Aug. doi: 10.1109/TPAMI.2012.231.

Fröhlich, B., E. Rodner, and J. Denzler. 2013. Semantic segmentation with millions of features: Integrating multiple cues in a combined random forest approach. In *Computer Vision – ACCV 2012*, 218–31. Berlin, Heidelberg: Springer Berlin Heidelberg.

Fu, J., J. Liu, Y. Wang, and H. Lu (2017). Stacked deconvolutional network for semantic segmentation. *CoRR abs/1708.04943*.

Ganin, Y., and V. S. Lempitsky (2014). N4-fields: Neural network nearest neighbor fields for image transforms. *CoRR abs/1406.6558*.

Garcia-Garcia, A., S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. G. Rodrguez (2017). A review on deep learning techniques applied to semantic segmentation. *CoRR abs/ 1704.06857*.

Geiger, A., P. Lenz, C. Stiller, and R. Urtasun. 2013. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research* 32(11):1231–37. September. doi: 10.1177/ 0278364913491297.

Ghiasi, G., T. Lin, and Q. V. Le (2019). Nas-fpn: Learning scalable feature pyramid architecture for object detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* Long Beach, CA, USA, 7029–38.

Girshick, R. B. (2015). Fast r-cnn. *CoRR*.

Girshick, R. B., J. Donahue, T. Darrell, and J. Malik (2013). Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR abs/1311.2524*.

Guo, Y., Y. Liu, T. Georgiou, and M. S. Lew. 2018. A review of semantic segmentation using deep neural networks. *International Journal of Multimedia Information Retrieval* 7(2):87–93. Jun. doi: 10.1007/s13735-017-0141-z.

Hariharan, B., P. Arbeláez, L. Bourdev, S. Maji, and J. Malik (2011, Nov). Semantic contours from inverse detectors. In *2011 International Conference on Computer Vision* Barcelona, Spain, 991–98.

Hariharan, B., P. Arbeláez, R. Girshick, and J. Malik. 2014. Simultaneous detection and segmentation. In *Computer Vision – ECCV 2014*, 297–312. Cham: Springer International Publishing.

He, K., G. Gkioxari, P. Dollár, and R. B. Girshick (2017). Mask r-cnn. *2017 IEEE International Conference on Computer Vision (ICCV)* Venice, 2980–88.

He, R., J. Yang, and X. Qi (2021). Re-distributing biased pseudo labels for semi-supervised semantic segmentation: A baseline investigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6930–40.

He, X., and R. S. Zemel. 2008. Learning hybrid models for image annotation with partially labeled data. In *Advances in neural information processing systems*, vol. 21, 625–32. Curran Associates, Inc.

He, K., X. Zhang, S. Ren, and J. Sun. 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37(9):1904–16. September. doi: 10.1109/TPAMI.2015.2389824.

He, K., X. Zhang, S. Ren, and J. Sun (2016, June). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* Las Vegas, NV, USA, 770–78.

Hong, W., Z. Wang, M. Yang, and J. Yuan (2018). Conditional generative adversarial network for structured domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* Salt Lake, UT, USA, 1335–44.

Howard, A., M. Sandler, B. Chen, W. Wang, L. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, and R. Pang, et al. (2019). Searching for mobilenetv3. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* Seoul, Korea, 1314–24.

Hsu, -C.-C., K.-J. Hsu, -C.-C. Tsai, -Y.-Y. Lin, and -Y.-Y. Chuang. 2019. Weakly supervised instance segmentation using the bounding box tightness prior. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vacouver CANADA.

Hu, P., F. Perazzi, F. C. Heilbron, O. Wang, Z. Lin, K. Saenko, and S. Sclaroff. 2020. Real-time semantic segmentation with fast attention. *IEEE Robotics and Automation Letters* 6 (1):263–70. doi:10.1109/LRA.2020.3039744.

Huang, G., S. Liu, L. Van der Maaten, and K. Q. Weinberger (2018). Condensenet: An efficient densenet using learned group convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* Salt Lake, UT, USA, 2752–61.

Huang, Y., Q. Wang, W. Jia, and X. He (2019). See more than once – Kernel-sharing atrous convolution for semantic segmentation.

Huang, G., J. Zhu, J. Li, Z. Wang, L. Cheng, L. Liu, H. Li, and J. Zhou. 2020. Channel-attention u-net: Channel attention mechanism for semantic segmentation of esophagus and esophageal cancer. *IEEE Access 8* 122798–810. doi:10.1109/ACCESS.2020.3007719.

Iglovikov, V., S. Seferbekov, A. Buslaev, and A. Shvets (2018). Ternausnetv2: Fully convolutional network for instance segmentation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* Salt Lake, UT, USA, 228–2284.

Jadon, S. (2020). A survey of loss functions for semantic segmentation. In *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational* Biology *(CIBCB)* Saint Louis, MO, USA, 1–7. IEEE.

Jiang, F., A. Grigorev, S. Rho, Z. Tian, Y. Fu, W. Jifara, A. Khan, and S. Liu. 2017. Medical image semantic segmentation based on deep learning. *Neural Computing & Applications*, 29 (5): 1257–1265 .

Jiang, P.-T., L.-H. Han, Q. Hou, -M.-M. Cheng, and Y. Wei. 2021. Online attention accumulation for weakly supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1–1. doi:10.1109/TPAMI.2021.3092573.

Jin, Z., T. Gong, D. Yu, Q. Chu, J. Wang, C. Wang, and J. Shao (2021). Mining contextual information beyond image for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7231–41.

Kang, G., Y. Wei, Y. Yang, Y. Zhuang, and A. G. Hauptmann (2020). Pixel-level cycle association: A new perspective for domain adaptive semantic segmentation. *arXiv preprint arXiv:2011.00147*.

Karimi, D., and S. E. Salcudean. 2019. Reducing the Hausdorff distance in medical image segmentation with convolutional neural networks. *IEEE Transactions on Medical Imaging* 39 (2):499–513. doi:10.1109/TMI.2019.2930068.

Ke, L., Y.-W. Tai, and C.-K. Tang (2021). Deep occlusion-aware instance segmentation with overlapping bilayers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4019–28.

Kemker, R., C. Salvaggio, and C. Kanan. 2018. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing* 145:60–77. doi:10.1016/j.isprsjprs.2018.04.014.

Khoreva, A., R. Benenson, J. Hosang, M. Hein, and B. Schiele (2017). Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* Venice, Italy, 876–85.

Kim, M., S. Woo, D. Kim, and I. S. Kweon (2021). The devil is in the boundary: Exploiting boundary representation for basis-based instance segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 929–38.

Kirillov, A., K. He, R. B. Girshick, C. Rother, and P. Dollar (2018). Panoptic segmentation. *CoRR abs/1801.00868*.

Kolesnikov, A., and C. H. Lampert (2016). Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European conference on computer vision* Amsterdam, the Netherlands, 695–711. Springer.

Krähenbühl, P., and V. Koltun. 2011. Efficient inference in fully connected crfs with Gaussian edge potentials. In *Advances in neural information processing systems*, Granada, Spain, 109–17.

Krizhevsky, A., I. Sutskever, and G. E. Hinton (2012a, 01). Imagenet classification with deep convolutional neural networks. *Neural Information Processing Systems* Lake Tahoe, Nevada, USA *25*.

Kweon, H., S.-H. Yoon, H. Kim, D. Park, and K.-J. Yoon (2021). Unlocking the potential of ordinary classifier: Class-specific adversarial erasing framework for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6994–7003.

Ladický, P., S. K. Alahari, C. Russell, and P. H. S. Torr. 2010. What, where and how many? Combining object detectors and crfs. In *Computer Vision – ECCV 2010*, ed. K. Daniilidis, P. Maragos, and N. Paragios, 424–37. Berlin, Heidelberg: Springer Berlin Heidelberg.

Ladický, L., C. Russell, P. Kohli, and P. H. S. Torr (2009, Sep.). Associative hierarchical CRFs for object class image segmentation. In *2009 IEEE 12th International Conference on Computer Vision* Kyoto, Japan, 739–46.

Lafferty, J., A. McCallum, and F. C. Pereira (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Lateef, F., and Y. Ruichek. 2019. Survey on semantic segmentation using deep learning techniques. *Neurocomputing* 338:321–48. doi:10.1016/j.neucom.2019.02.003.

Lazebnik, S., C. Schmid, and J. Ponce (2006, June). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR06)* New York, NY, USA, Volume 2, 2169–78.

LeCun, Y. . 1989. Generalization and network design strategies. In *Connectionism in perspective*, Zurich, Switzerland: Elsevier, 143–55.

Lee, Y., and J. Park (2020). Centermask: Real-time anchor-free instance segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13903–12.

Lee, H., F. M. Troschel, S. Tajmir, G. Fuchs, J. Mario, F. J. Fintelmann, and S. Do. 2017. Pixel-level deep segmentation: Artificial intelligence quantifies muscle on computed tomography for body morphometric analysis. *Journal of Digital Imaging* 30(4):487–98. Aug. doi: 10.1007/s10278-017-9988-z.

Lee, J., J. Yi, C. Shin, and S. Yoon (2021). Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2643–52.

Lempitsky, V., A. Vedaldi, and A. Zisserman. 2011. Pylon model for semantic segmentation. In *Advances in neural information processing systems*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger. ed., vol. 24. 1485–93. Granada, Spain: Curran Associates, Inc.

Li, Y., Z. Kuang, L. Liu, Y. Chen, and W. Zhang (2021b). Pseudo-mask matters in weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6964–73.

Li, Z., Y. Sun, L. Zhang, and J. Tang. 2021c. Ctnet: Context-based tandem network for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1–1. doi:10.1109/TPAMI.2021.3132068.

Li, H., P. Xiong, J. An, and L. Wang (2018). Pyramid attention network for semantic segmentation. In *British Machine Vision Conference 2018, BMVC 2018*, *Newcastle, UK*, September 3-6, 2018, 285. BMVA Press.

Li, H., P. Xiong, H. Fan, and J. Sun (2019a). Dfanet: Deep feature aggregation for real-time semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 9522–31.

Li, R., S. Zheng, C. Zhang, C. Duan, L. Wang, and P. M. Atkinson. 2021a. Abcnet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery. *ISPRS Journal of Photogrammetry and Remote Sensing* 181:84–98. doi:10.1016/j.isprsjprs.2021.09.005.

Li, X., Z. Zhong, J. Wu, Y. Yang, Z. Lin, and H. Liu (2019b). Expectation-maximization attention networks for semantic segmentation.

Liang, X., X. Shen, J. Feng, L. Lin, and S. Yan. 2016. Semantic object parsing with graph lstm. In *Computer Vision – ECCV 2016*, ed. B. Leibe, J. Matas, N. Sebe, and M. Welling, Amsterdam, The Netherlands: Springer International Publishing, 125–43.

Lin, D., J. Dai, J. Jia, K. He, and J. Sun (2016a). Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, Las Vegas, NV, USA, 3159–67.

Lin, D., Y. Ji, D. Lischinski, D. Cohen-Or, and H. Huang (2018, September). Multi-scale context intertwining for semantic segmentation. In *The European Conference on Computer Vision (ECCV)*, Munich, Germany.

Lin, T.-Y., M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–55. Zurich, Switzerland: Springer.

Lin, G., C. Shen, A. V. D. Hengel, and I. Reid (2016b, June). Efficient piecewise training of deep structured models for semantic segmentation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 3194–203.

Liu, W., D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg (2016). SSD: Single shot multibox detector. In *Computer Vision - ECCV 2016-14th European Conference*, Amsterdam, The Netherlands, 21–37.

Liu, Z., X. Li, P. Luo, C. Loy, and X. Tang (2015, Dec). Semantic image segmentation via deep parsing network. In *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 1377–85.

Liu, W., A. Rabinovich, and A. C. Berg (2015). Parsenet: Looking wider to see better.

Liu, Y., Y.-H. Wu, P.-S. Wen, Y.-J. Shi, Y. Qiu, and -M.-M. Cheng. 2020. Leveraging instance-, image-and dataset-level information for weakly supervised instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1–1. doi:10.1109/TPAMI.2020.3023152.

Liu, Y., W. Zhang, and J. Wang (2021). Source-free domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1215–24.

Long, J., E. Shelhamer, and T. Darrell (2014). Fully convolutional networks for semantic segmentation. *CoRR abs/1411.4038*.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60(2):91–110. Nov. doi: 10.1023/B:VISI.0000029664.99615.94.

Lu, H., L. Fang, M. Lin, and Z. Deng (2021). Feature enhanced projection network for zero-shot semantic segmentation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*,Xi'an, China, 14011–17. IEEE.

Luo, P., G. Wang, L. Lin, and X. Wang (2017, Oct). Deep dual learning for semantic image segmentation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2737–45.

Marin, D., Z. He, P. Vajda, P. Chatterjee, S. Tsai, F. Yang, and Y. Boykov (2019). Efficient segmentation: Learning downsampling near semantic boundaries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, Korea, 2131–41.

Mičušĺík, B., and J. Košecká (2009, Sep.). Semantic segmentation of street scenes by superpixel co-occurrence and 3d geometry. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, Kyoto, Japan, 625–32.

Minaee, S., Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos (2020). Image segmentation using deep learning: A survey.

Montillo, A., J. Shotton, J. Winn, J. E. Iglesias, D. Metaxas, and A. Criminisi. 2011. Entangled decision forests and their application for semantic segmentation of ct images. In *Information processing in medical imaging*, Gábor Székely and Horst K. Hahn editors, 184–96. Berlin, Heidelberg: Springer Berlin Heidelberg.

Mottaghi, R., X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille (2014). The role of context for object detection and semantic segmentation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, OH, USA.

Ning, F., D. Delhomme, Y. LeCun, F. Piano, L. Bottou, and P. E. Barbano. 2005. Toward automatic phenotyping of developing embryos from videos. *IEEE Transactions on Image Processing* 14(9):1360–71. Sept. doi: 10.1109/TIP.2005.852470.

Noh, H., S. Hong, and B. Han (2015). Learning deconvolution network for semantic segmentation. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, Washington, DC, USA, 1520–28. IEEE Computer Society.

Oktay, O., J. Schlemper, L. L. Folgoc, M. C. H. Lee, M. P. Heinrich, K. Misawa, K. Mori, S. G. McDonagh, N. Y. Hammerla, B. Kainz, et al. (2018). Attention u-net: Learning where to look for the pancreas. *CoRR abs/1804.03999*.

Orsic, M., I. Kreso, P. Bevandic, and S. Segvic (2019). In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 12607–16.

Oršić, M., and S. Šegvić. 2021. Efficient semantic segmentation with pyramidal fusion. *Pattern Recognition* 110:107611. doi:10.1016/j.patcog.2020.107611.

Pan, F., I. Shin, F. Rameau, S. Lee, and I. S. Kweon (2020). Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3764–73.

Pastore, G., F. Cermelli, Y. Xian, M. Mancini, Z. Akata, and B. Caputo (2021). A closer look at self-training for zero-label semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2693–702.

Pathak, D., P. Krahenbuhl, and T. Darrell (2015). Constrained convolutional neural networks for weakly supervised segmentation. In *Proceedings of the IEEE international conference on computer vision*, Santiago, Chile, 1796–804.

Peng, S., W. Jiang, H. Pi, X. Li, H. Bao, and X. Zhou (2020). Deep snake for real-time instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8533–42.

Peng, C., X. Zhang, G. Yu, G. Luo, and J. Sun (2017, July). Large kernel matters — Improve semantic segmentation by global convolutional network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 1743–51.

Perreault, H., G.-A. Bilodeau, N. Saunier, and M. Héritier (2021). Centerpoly: Real-time instance segmentation using bounding polygons. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2982–91.

Pfeuffer, A., K. Schulz, and K. Dietmayer (2019). Semantic segmentation of video sequences with convolutional lstms. *CoRR abs/1905.01058*.

Pinheiro, P. O., and R. Collobert (2014). Recurrent convolutional neural networks for scene labeling. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, Beijing, China, I–82–I–90. JMLR.org.

Pinheiro, P. O., and R. Collobert (2015). From image-level to pixel-level labeling with convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, Boston, MA, USA, 1713–21.

Pinheiro, P. O., R. Collobert, and P. Dollar. 2015. Learning to segment object candidates. In *Advances in neural information processing systems*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett. ed. vol. 28, 1990–98. Montreal, Canada: Curran Associates, Inc.

Pinheiro, P. O., T.-Y. Lin, R. Collobert, and P. Dollár. 2016. Learning to refine object segments. In *Computer Vision – ECCV 2016*, ed. B. Leibe, J. Matas, N. Sebe, and M. Welling, 75–91. Cham: Springer International Publishing.

Prest, A., C. Leistner, J. Civera, C. Schmid, and V. Ferrari (2012, June). Learning object class detectors from weakly annotated video. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, 3282–89.

Rav, D., M. Bober, G. Farinella, M. Guarnera, and S. Battiato. 2016. Semantic segmentation of images exploiting dct based features and random forest. *Pattern Recognition* 52(C):260–73. April. doi: 10.1016/j.patcog.2015.10.021.

Redmon, J., S. Divvala, R. Girshick, and A. Farhadi (2016, Jun). You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA, 779–88. IEEE Computer Society.

Ren, S., K. He, R. Girshick, and J. Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, ed., C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, vol. 28, NY, United States: Curran Associates Inc., 91–99.

Richter, S. R., V. Vineet, S. Roth, and V. Koltun (2016). Playing for data: Ground truth from computer games. In *European conference on computer vision*, Amsterdam, The Netherlands, 102–18. Springer.

Ronneberger, O., P. Fischer, and T. Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Navab, Nassir and Hornegger, Joachi and Wells, William M.and Frangi, Alejandro F. editors, 234–41. Cham: Springer International Publishing.

Ros, G., L. Sellart, J. Materzynska, D. Vazquez, and A. Lopez (2016a). The SYNTHIA Dataset: A large collection of synthetic images for semantic segmentation of urban scenes.

Ros, G., L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez (2016b). The Synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, Las Vegas, NV, USA, 3234–43.

Rosas-Arias, L., G. Benitez-Garcia, J. Portillo-Portillo, G. Sánchez-Pérez, and K. Yanai (2021). Fast and accurate real-time semantic segmentation with dilated asymmetric convolutions. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 2264–71. IEEE, Milan, Italy.

Rosenholtz, R. 2016. Capabilities and limitations of peripheral vision. *Annual Review of Vision Science* 2 (1):437–57. doi:10.1146/annurev-vision-082114-035733.

Rother, C., V. Kolmogorov, and A. Blake. 2004. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics (TOG)* 23(ACM):309–14. doi:10.1145/1015706.1015720.

Saffar, M. H., M. Fayyaz, M. Sabokrou, and M. Fathy (2018). Semantic video segmentation: A review on recent approaches.

Saha, M., and C. Chakraborty. 2018. Her2net: A deep framework for semantic segmentation and classification of cell membranes and nuclei in breast cancer evaluation. *IEEE Transactions on Image Processing* 27(5):2189–200. May. doi: 10.1109/TIP.2018.2795742.

Saporta, A., T.-H. Vu, M. Cord, and P. Pérez (2021, October). Multi-target adversarial frameworks for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9072–81.

Shelhamer, E., J. Long, and T. Darrell. 2017. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (4):640–51. April. doi: 10.1109/TPAMI.2016.2572683.

Shen, Y., L. Cao, Z. Chen, F. Lian, B. Zhang, C. Su, Y. Wu, F. Huang, and R. Ji (2021a). Toward joint thing-and-stuff mining for weakly supervised panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16694–705.

Shen, Y., L. Cao, Z. Chen, B. Zhang, C. Su, Y. Wu, F. Huang, and R. Ji (2021b). Parallel detection-and-segmentation learning for weakly supervised instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8198–208.

Shi, H., H. Li, F. Meng, Q. Wu, L. Xu, and K. N. Ngan. 2018. Hierarchical parsing net: Semantic scene parsing from global scene to objects. *IEEE Transactions on Multimedia* 20 (10):2670–82. Oct. doi: 10.1109/TMM.2018.2812600.

Shin, I., D.-J. Kim, J. W. Cho, S. Woo, K. Park, and I. S. Kweon (2021). Labor: Labeling only if required for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8588–98.

Shotton, J., M. Johnson, and R. Cipolla (2008, June). Semantic texton forests for image categorization and segmentation. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 1–8.

Shuai, B., Z. Zuo, B. Wang, and G. Wang (2016, June). Dag-recurrent neural networks for scene labeling. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 3620–29.

Siam, M., S. Elkerdawy, M. Jägersand, and S. Yogamani (2017). Deep semantic segmentation for automated driving: Taxonomy, roadmap and challenges. In *20th IEEE International Conference on Intelligent Transportation Systems, ITSC 2017*, *Yokohama, Japan*, October 16-19, 2017, 1–8.

Simonyan, K., and A. Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proceedings of Workshop at International Conference on Learning Representations (ICLR) Workshops*, San Diego, CA, USA.

Sun, J. (2019). Mask-yolo: Efficient instance-level segmentation network based on yolo-v2. *GitHub Repository*.

Sun, K., H. Shi, Z. Zhang, and Y. Huang (2021). Ecs-net: Improving weakly supervised semantic segmentation by using connections between class activation maps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7283–92.

Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. 2015. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, pp. 1-9, doi: 10.1109/CVPR.2015.7298594.

Teichmann, M. T. T., and R. Cipolla (2018). Convolutional CRFs for semantic segmentation. *CoRR abs/1805.04777*.

Teja Sivaprasad, Prabhu, and Fleuret, F., et al. (2021). Uncertainty reduction for model adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9613–23.

Thoma, M. (2016). A survey of semantic segmentation. *arXiv preprint arXiv:1602.06541*.

Tian, Z., C. Shen, X. Wang, and H. Chen (2021). Boxinst: High-performance instance segmentation with box annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5443–52.

Tighe, J., and S. Lazebnik (2010). Superparsing: Scalable nonparametric image parsing with superpixels. In *European Conference on Computer Vision – ECCV 2010*, Heraklion, Crete, Greece, 352–65.

Ulku, I., P. Barmpoutis, T. Stathaki, and E. Akagündüz (2019). Comparison of single channel indices for u-net-based segmentation of vegetation in satellite images. In *12th International Conference on Machine Vision (ICMV19)*, SPIE Proceedings, Amsterdam, The Netherlands.

Ulusoy, I., and C. M. Bishop (2005). Generative versus discriminative methods for object recognition. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, 20-26 June 2005, *San Diego, CA, USA*, 258–65.

Varma, G., A. Subramanian, A. M. Namboodiri, M. Chandraker, and C. V. Jawahar (2018). IDD: A dataset for exploring problems of autonomous navigation in unconstrained environments. *CoRR abs/1811.10200*.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin (2017). Attention is all you need. *CoRR abs/1706.03762*.

Vezhnevets, A., V. Ferrari, and J. M. Buhmann (2011, Nov). Weakly supervised semantic segmentation with a multi-image model. In *2011 International Conference on Computer Vision*, Barcelona, Spain, 643–50.

Visin, F., A. Romero, K. Cho, M. Matteucci, M. Ciccone, K. Kastner, Y. Bengio, and A. Courville (2016, June). Reseg: A recurrent neural network-based model for semantic segmentation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Las Vegas, NV, USA, 426–33.

Vu, T.-H., H. Jain, M. Bucher, M. Cord, and P. Pérez (2019). Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2517–26.

Wang, Q., D. Dai, L. Hoyer, O. Fink, and L. Van Gool (2021). Domain adaptive semantic segmentation with self-supervised depth estimation. *arXiv preprint arXiv:2104.13613*.

Wang, D., G. Hu, and C. Lyu. 2020. Frnet: An end-to-end feature refinement neural network for medical image segmentation. *The Visual Computer* 37:1113–27. May. doi: 10.1007/s00371-020-01856-y.

Wang, X., T. Kong, C. Shen, Y. Jiang, and L. Li (2019b). Solo: Segmenting objects by locations. In European Conference on Computer Vision (pp. 649-665). Springer, Cham.

Wang, K., J. H. Liew, Y. Zou, D. Zhou, and J. Feng (2019a). Panet: Few-shot image semantic segmentation with prototype alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, Korea, 9197–206.

Wang, Y., J. Peng, and Z. Zhang (2021). Uncertainty-aware pseudo label refinery for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9092–101.

Wang, P., X. Shen, Z. Lin, S. Cohen, B. Price, and A. Yuille (2015, Dec). Joint object and part segmentation using deep learned potentials. In *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 1573–81.

Wang, Y., J. Zhang, M. Kan, S. Shan, and X. Chen (2020b). Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12275–84.

Wang, X., R. Zhang, T. Kong, L. Li, and C. Shen (2020a). Solov2: Dynamic and fast instance segmentation. *arXiv preprint arXiv:2003.10152*.

Wu, X., Z. Wu, H. Guo, L. Ju, and S. Wang (2021). Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15769–78.

Xia, X., Q. Lu, and X. Gu. 2019. Exploring an easy way for imbalanced data sets in semantic image segmentation. *Journal of Physics. Conference Series* 1213(2):022003. Jun.

Xian, Y., S. Choudhury, Y. He, B. Schiele, and Z. Akata (2019). Semantic projection network for zero-and few-label semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 8256–65.

Xiao, J., and L. Quan (2009, Sep.). Multiple view semantic segmentation for street view images. In 2009 IEEE 12th international conference on computer vision (pp. 686-693). IEEE.

Xie, G.-S., J. Liu, H. Xiong, and L. Shao (2021). Scale-aware graph neural network for few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5475–84.

Xie, Q., M. T. Luong, E. Hovy, and Q. V. Le (2020b). Self-training with noisy student improves imagenet classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–95.

Xie, E., P. Sun, X. Song, W. Wang, X. Liu, D. Liang, C. Shen, and P. Luo (2020a). Polarmask: Single shot instance segmentation with polar representation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12190–99.

Xu, L., W. Ouyang, M. Bennamoun, F. Boussaid, F. Sohel, and 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*, 6984–93.

Xu, J., A. G. Schwing, and R. Urtasun (2015). Learning to segment under various forms of weak supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, Boston, MA, USA, 3781–90.

Xu, W., H. Wang, F. Qi, and C. Lu (2019a). Explicit shape encoding for real-time instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, Korea, 5168–77.

Xu, W., H. Wang, F. Qi, and C. Lu (2019b). Explicit shape encoding for real-time instance segmentation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea, 5167–76.

Yang, Y., S. Hallman, D. Ramanan, and C. C. Fowlkes. 2012. Layered object models for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34 (9):1731–43. doi:10.1109/TPAMI.2011.208.

Yao, Y., T. Chen, G.-S. Xie, C. Zhang, F. Shen, Q. Wu, Z. Tang, and J. Zhang (2021). Non-salient region object mining for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2623–32.

Yao, J., S. Fidler, and R. Urtasun (2012, June). Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, 702–09.

Yu, F., and V. Koltun (2015). Multi-scale context aggregation by dilated convolutions. *CoRR abs/1511.07122*.

Yu, C., J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang (2018a). Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, Munich Germany, 325–41.

Yu, C., J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang (2018b, June). Learning a discriminative feature network for semantic segmentation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 1857–66.

Yu, H., Y. Zhengeng, L. Tan, Y. Wang, W. Sun, M. Sun, and Y. Tang. 2018 05. Methods and datasets on semantic segmentation: A review. *Neurocomputing* 304, 82-103.

Yuan, Y., X. Chen, and J. Wang (2020). Object-contextual representations for semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference*, *Glasgow, UK*, August 23–28,2020, *Proceedings, Part VI 16*, 173–90. Springer.

Zaitoun, N. M., and M. J. Aqel (2015). Survey on image segmentation techniques. *Procedia Computer Science* 65, 797–806. International Conference on Communications, management, and Information technology (ICCMIT'2015), Prague, Czech Republic.

Zhang, H., K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal (2018a, June). Context encoding for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA.

Zhang, R., Z. Tian, C. Shen, M. You, and Y. Yan (2020a). Mask encoding for single shot instance segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10223–32.

Zhang, F., P. Torr, R. Ranftl, and S. Richter. 2021. Looking beyond single images for contrastive semantic segmentation learning. In *Advances in neural information processing systems*, 34.

Zhang, X., H. Xu, H. Mo, J. Tan, C. Yang, and W. Ren (2020b). Dcnas: Densely connected neural architecture search for semantic image segmentation. *CoRR*.

Zhang, Z., X. Zhang, C. Peng, X. Xue, and J. Sun. 2018c. Exfuse: Enhancing feature fusion for semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, (pp. 269-284).

Zhang, H., H. Zhang, C. Wang, and J. Xie (2019, June). Co-occurrent features in semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA.

Zhang, X., X. Zhou, M. Lin, and J. Sun (2018b). Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, Salt Lake City, UT, USA, 6848–56.

Zhao, S., B. Li, X. Yue, Y. Gu, P. Xu, R. Hu, H. Chai, and K. Keutzer (2019). Multi-source domain adaptation for semantic segmentation. *arXiv preprint arXiv:1910.12181*.

Zhao, H., X. Qi, X. Shen, J. Shi, and J. Jia (2018a). ICNet for real-time semantic segmentation on high-resolution images. In *ECCV*, Munich, Germany.

Zhao, H., J. Shi, X. Qi, X. Wang, and J. Jia (2017, July). Pyramid scene parsing network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 6230–39.

Zhao, X., R. Vemulapalli, P. A. Mansfield, B. Gong, B. Green, L. Shapira, and Y. Wu (2021). Contrastive learning for label efficient semantic segmentation. In *Proceedings of the IEEE/ CVF International Conference on Computer Vision*, 10623–33.

Zhao, H., Y. Zhang, S. Liu, J. Shi, C. C. Loy, D. Lin, and J. Jia. 2018b. Psanet: Point-wise spatial attention network for scene parsing. In *Computer Vision – ECCV 2018*, ed. V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, 270–86. Cham: Springer International Publishing.

Zheng, S., S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr (2015, Dec). Conditional random fields as recurrent neural networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 1529–37.

Zheng, Z., and Y. Yang. 2021. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision* 129 (4):1106–20. doi:10.1007/s11263-020-01395-y.

Zhong, Z., J. Li, W. Cui, and H. Jiang (2016, July). Fully convolutional networks for building and road extraction: Preliminary results. In *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Beijing, China, 1591–94.

Zhong, G., Y.-H. Tsai, and M.-H. Yang (2016b). Weakly-supervised video scene co-parsing. In *Asian Conference on Computer Vision*, 20–36. Taipei, Taiwan: Springer.

Zhou, B., A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, Las Vegas, NV, USA, 2921–29.

Zhou, B., H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba. 2019. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision* 127 (3):302–21. doi:10.1007/s11263-018-1140-0.

Zoph, B., G. Ghiasi, T. Lin, Y. Cui, H. Liu, E. D. Cubuk, and Q. Le (2020). Rethinking pre-training and self-training. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, December 6- 12,2020, *virtual.*

Zoph, B., V. Vasudevan, J. Shlens, and Q. V. Le (2018). Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, Salt Lake City, UT, USA, 8697–710.

Zuo, Y., and T. Drummond (2017, 13–15 Nov). Fast residual forests: Rapid ensemble learning for semantic segmentation. In S. Levine, V. Vanhoucke, and K. Goldberg (Eds.), *Proceedings of the 1st Annual Conference on Robot Learning*, Volume 78 of *Proceedings of Machine Learning Research*, Mountain View, California, USA, 27–36. PMLR.