Image Segmentation Using Deep Learning: A Survey

Shervin Minaee, *Member, IEEE,* Yuri Boykov, *Member, IEEE,* Fatih Porikli, *Fellow, IEEE,* Antonio Plaza, *Fellow, IEEE,* Nasser Kehtarnavaz, *Fellow, IEEE,* and Demetri Terzopoulos, *Fellow, IEEE*

Abstract—Image segmentation is a key task in computer vision and image processing with important applications such as scene understanding, medical image analysis, robotic perception, video surveillance, augmented reality, and image compression, among others, and numerous segmentation algorithms are found in the literature. Against this backdrop, the broad success of Deep Learning (DL) has prompted the development of new image segmentation approaches leveraging DL models. We provide a comprehensive review of this recent literature, covering the spectrum of pioneering efforts in semantic and instance segmentation, including convolutional pixel-labeling networks, encoder-decoder architectures, multiscale and pyramid-based approaches, recurrent networks, visual attention models, and generative models in adversarial settings. We investigate the relationships, strengths, and challenges of these DL-based segmentation models, examine the widely used datasets, compare performances, and discuss promising research directions.

Index Terms—Image segmentation, deep learning, convolutional neural networks, encoder-decoder models, recurrent models, generative models, semantic segmentation, instance segmentation, panoptic segmentation, medical image segmentation.

1 Introduction

Image segmentation has been a fundamental problem in computer vision since the early days of the field [1] (Chapter 8). An essential component of many visual understanding systems, it involves partitioning images (or video frames) into multiple segments and objects [2] (Chapter 5) and plays a central role in a broad range of applications [3] (Part VI), including medical image analysis (e.g., tumor boundary extraction and measurement of tissue volumes), autonomous vehicles (e.g., navigable surface and pedestrian detection), video surveillance, and augmented reality to name a few.

Image segmentation can be formulated as the problem of classifying pixels with semantic labels (semantic segmentation), or partitioning of individual objects (instance segmentation), or both (panoptic segmentation). Semantic segmentation performs pixel-level labeling with a set of object categories (e.g., human, car, tree, sky) for all image pixels; thus, it is generally a more demanding undertaking than whole-image classification, which predicts a single label for the entire image. Instance segmentation extends the scope of semantic segmentation by detecting and delineating each object of interest in the image (e.g., individual people).

Numerous image segmentation algorithms have been developed in the literature, from the earliest methods, such as thresholding [4], histogram-based bundling, region-growing [5], k-means clustering [6], watershed methods [7], to more advanced algorithms such as active contours [8], graph cuts [9], conditional and Markov random fields [10], and sparsity-based [11], [12] methods. In recent years, however, deep learning (DL) models have yielded a new generation of image segmentation models with remarkable performance improvements, often achieving the highest accuracy rates on popular benchmarks (e.g., Fig. 1). This has caused a paradigm shift in the field.

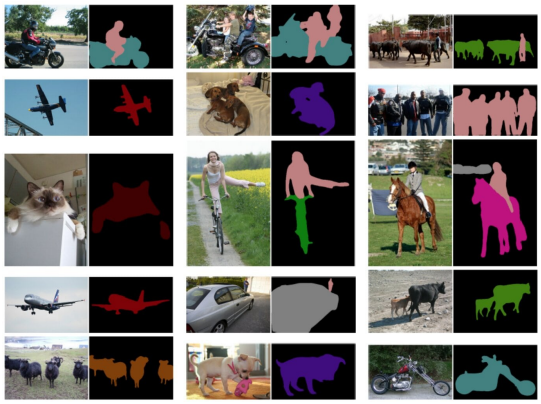


Fig. 1. Segmentation results of DeepLabV3 [13] on sample images.

This survey, a revised version of [14], covers the recent literature in deep-learning-based image segmentation, including more than 100 such segmentation methods proposed to date. It provides a comprehensive review with insights into different aspects of these methods, including the training data, the choice of network architectures, loss functions, training strategies, and their key contributions. The target literature is organized into the following categories:

1) Fully convolutional networks
2) Convolutional models with graphical models
3) Encoder-decoder based models
4) Multiscale and pyramid network based models
5) R-CNN based models (for instance segmentation)
6) Dilated convolutional models and DeepLab family
7) Recurrent neural network based models

- *S. Minaee is with Snapchat Machine Learning Research.*
- *Y. Boykov is with the University of Waterloo.*
- *F. Porikli is with the Australian National University, and Huawei.*
- *A. Plaza is with the University of Extremadura, Spain.*
- *N. Kehtarnavaz is with the University of Texas at Dallas.*
- *D. Terzopoulos is with the University of California, Los Angeles.*

Manuscript received December ??, 2019; revised ?? ??, 2021.

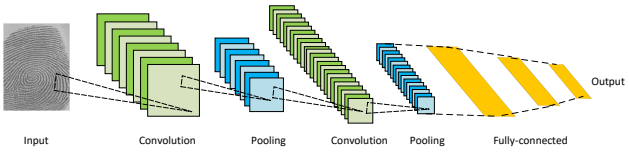


Fig. 2. Architecture of CNNs. From [15].

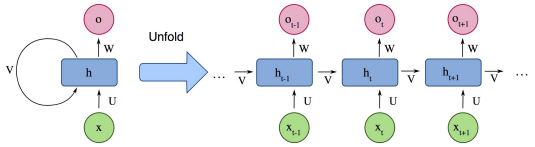


Fig. 3. Architecture of a simple RNN. Courtesy of Christopher Olah [21].

8) Attention-based models
9) Generative models and adversarial training
10) Convolutional models with active contour models
11) Other models

Within this taxonomy,

- we provide a comprehensive review and analysis of deep-learning-based image segmentation algorithms;
- we overview popular image segmentation datasets, grouped into 2D and 2.5D (RGB-D) images;
- we summarize the performances of the reviewed segmentation methods on popular benchmarks;
- we discuss several challenges and future research directions for deep-learning-based image segmentation.

The remainder of this survey is organized as follows: Section 2 overviews popular Deep Neural Network (DNN) architectures that serve as the backbones of many modern segmentation algorithms. Section 3 reviews the most significant state-of-the-art deep learning based segmentation models. Section 4 overviews some of the most popular image segmentation datasets and their characteristics. Section 5 lists popular metrics for evaluating deep-learning-based segmentation models and tabulates model performances. Section 6 discusses the main challenges and opportunities of deep learning-based segmentation methods. Section 7 presents our conclusions.

2 Deep Neural Network Architectures

This section provides an overview of prominent DNN architectures used by the computer vision community, including convolutional neural networks, recurrent neural networks and long short-term memory, encoder-decoder and autoencoder models, and generative adversarial networks. Due to space limitations, several other DNN architectures that have been proposed, among them transformers, capsule networks, gated recurrent units, and spatial transformer networks, will not be covered.

2.1 Convolutional Neural Networks (CNNs)

CNNs are among the most successful and widely used architectures in the deep learning community, especially for computer vision tasks. CNNs were initially proposed by Fukushima [16] in his seminal paper on the "Neocognitron", which was based on Hubel and Wiesel's hierarchical receptive field model of the visual cortex. Subsequently, Waibel *et al.* [17] introduced CNNs with weights shared among temporal receptive fields and backpropagation training for phoneme recognition, and LeCun *et al.* [15] developed a practical CNN architecture for document recognition (Fig. 2). CNNs usually include three types of layers: i) convolutional layers, where a kernel (or filter) of weights is convolved to extract features; ii) nonlinear layers, which apply (usually element-wise) an activation function to feature maps, thus enabling the network to model nonlinear functions; and iii) pooling layers, which reduce spatial resolution by replacing small neighborhoods in a feature map with some statistical information about those neighborhoods (mean, max, etc.). The neuronal units in layers are locally connected; that is, each unit receives weighted inputs from a small neighborhood, known as the receptive field, of units in the previous layer. By stacking layers to form multiresolution pyramids, the higher-level layers learn features from increasingly wider receptive fields. The main computational advantage of CNNs is that all the receptive fields in a layer share weights, resulting in a significantly smaller number of parameters than fully-connected neural networks. Some of the most well known CNN architectures include AlexNet [18], VGGNet [19], and ResNet [20].

2.2 Recurrent Neural Networks (RNNs) and the LSTM

RNNs [22] are commonly used to process sequential data, such as speech, text, videos, and time-series. Referring to Fig. 3, at each time step t the model collects the input x_t and the hidden state h_{t-1} from the previous step, and outputs a target value o_t and the next hidden state h_{t+1}. RNNs are typically problematic for long sequences as they cannot capture long-term dependencies in many real-world applications and often suffer from gradient vanishing or exploding problems. However, a type of RNN known as the Long Short-Term Memory (LSTM) [23] is designed to avoid these issues. The LSTM architecture (Fig. 4) includes three gates (input gate, output gate, and forget gate) that regulate the flow of information into and out of a memory cell that stores values over arbitrary time intervals.

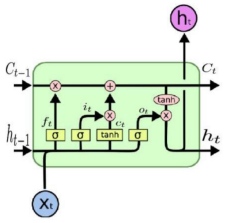


Fig. 4. Architecture of a standard LSTM module. Courtesy of Olah [21].

2.3 Encoder-Decoder and Auto-Encoder Models

Encoder-decoders [24], [25] are a family of models that learn to map data-points from an input domain to an output domain via a two-stage network (Fig. 5): The encoder,

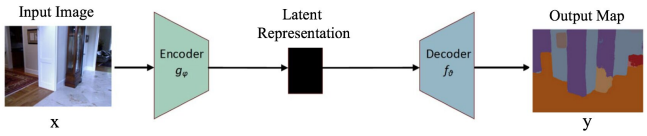


Fig. 5. Architecture of a simple encoder-decoder model.

performing an encoding function $z = g(x)$, compresses the input x into a latent-space representation z, while the decoder $y = f(z)$ predicts the output y from z. The latent, or feature (vector), representation captures the semantic information of the input useful in predicting the output. Such models are popular for sequence-to-sequence modeling in Natural Language Processing (NLP) applications as well as in image-to-image translation, where the output could be an enhanced version of the image (such as in image de-blurring, or super-resolution) or a segmentation map. Auto-encoders are a special case of encoder-decoder models in which the input and output are the same.

2.4 Generative Adversarial Networks (GANs)

GANs [26] are a newer family of deep learning models. They consist of two networks—a generator and a discriminator (Fig. 6). In the conventional GAN, the generator network G learns a mapping from noise z (with a prior distribution) to a target distribution y, which is similar to the "real" samples. The discriminator network D attempts to distinguish the generated "fake" samples from the real ones. The GAN may be characterized as a minimax game between G and D, where D tries to minimize its classification error in distinguishing fake samples from real ones, hence maximizing a loss function, and G tries to maximize the discriminator network's error, hence minimizing the loss function. GAN variants include Convolutional-GANs [27], conditional-GANs [28], and Wasserstein-GANs [29].

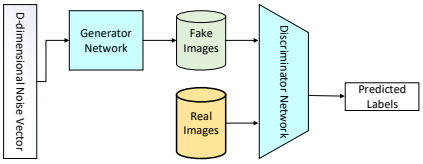


Fig. 6. Architecture of a GAN. Courtesy of Ian Goodfellow.

3 DL-Based Image Segmentation Models

This section is a survey of numerous learning-based segmentation methods, grouped into 10 categories based on their model architectures. Several architectural features are common among many of these methods, such as encoders and decoders, skip-connections, multiscale architectures, and more recently the use of dilated convolutions. It is convenient to group models based on their architectural contributions over prior models.

3.1 Fully Convolutional Models

Long *et al.* [30] proposed Fully Convolutional Networks (FCNs), a milestone in DL-based semantic image segmentation models. An FCN (Fig. 7) includes only convolutional layers, which enables it to output a segmentation map whose size is the same as that of the input image. To handle arbitrarily-sized images, the authors modified existing CNN architectures, such as VGG16 and GoogLeNet, by removing all fully-connected layers such that the model outputs a spatial segmentation map instead of classification scores.

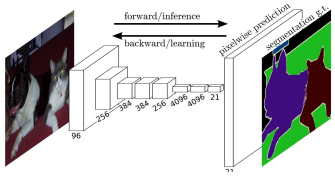


Fig. 7. The FCN learns to make pixel-accurate predictions. From [30].

Through the use of skip connections (Fig. 8) in which feature maps from the final layers of the model are upsampled and fused with feature maps of earlier layers, the model combines semantic information (from deep, coarse layers) and appearance information (from shallow, fine layers) in order to produce accurate and detailed segmentations. Tested on PASCAL VOC, NYUDv2, and SIFT Flow, the model achieved state-of-the-art segmentation performance.

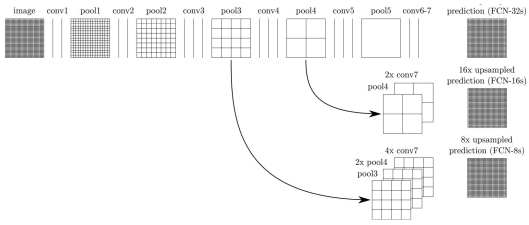


Fig. 8. Skip connections combine coarse and fine information. From [30].

FCNs have been applied to a variety of segmentation problems, such as brain tumor segmentation [31], instance-aware semantic segmentation [32], skin lesion segmentation [33], and iris segmentation [34]. While demonstrating that DNNs can be trained to perform semantic segmentation in an end-to-end manner on variable-sized images, the conventional FCN model has some limitations—it is too computationally expensive for real-time inference, it does not account for global context information in an efficient manner, and it is not easily generalizable to 3D images. Several researchers have attempted to overcome some of the limitations of the FCN. For example, Liu *et al.* [35] proposed ParseNet (Fig. 9), which adds global context to FCNs by using the average feature for a layer to augment the features at each location. The feature map for a layer is pooled over the whole image, resulting in a context vector. The context vector is normalized and unpooled to produce new feature maps of the same size as the initial ones, which are then concatenated, which amounts to an FCN whose convolutional layers are replaced by the described module (Fig. 9e).

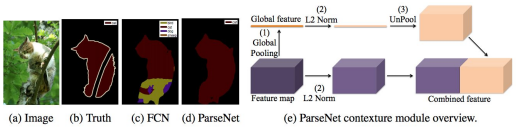


Fig. 9. The ParseNet (e) uses extra global context to produce a segmentation (d) smoother than that of an FCN (c). From [35].

3.2 CNNs With Graphical Models

As discussed, the FCN ignores potentially useful scene-level semantic context. To exploit more context, several approaches incorporate into DL architectures probabilistic graphical models, such as Conditional Random Fields (CRFs) and Markov Random Fields (MRFs).

Due to the invariance properties that make CNNs good for high level tasks such as classification, responses from the later layers of deep CNNs are not sufficiently well localized for accurate object segmentation. To address this drawback, Chen *et al.* [36] proposed a semantic segmentation algorithm that combines CNNs and fully-connected CRFs (Fig. 10). They showed that their model can localize segment boundaries with higher accuracy than was possible with previous methods.

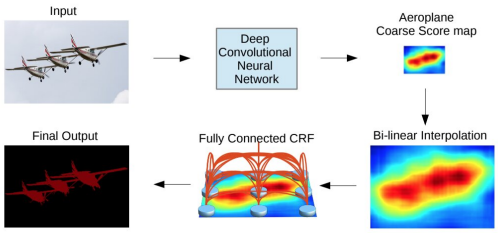


Fig. 10. A CNN+CRF model. From [36].

Schwing and Urtasun [37] proposed a fully-connected deep structured network for image segmentation. They jointly trained CNNs and fully-connected CRFs for semantic image segmentation, and achieved encouraging results on the challenging PASCAL VOC 2012 dataset. Zheng *et al.* [38] proposed a similar semantic segmentation approach. In related work, Lin *et al.* [39] proposed an efficient semantic segmentation model based on contextual deep CRFs. They explored "patch-patch" context (between image regions) and "patch-background" context to improve semantic segmentation through the use of contextual information.

Liu *et al.* [40] proposed a semantic segmentation algorithm that incorporates rich information into MRFs, including high-order relations and mixture of label contexts. Unlike previous efforts that optimized MRFs using iterative algorithms, they proposed a CNN model, namely a Parsing Network, which enables deterministic end-to-end computation in one pass.

3.3 Encoder-Decoder Based Models

Most of the popular DL-based segmentation models use some kind of encoder-decoder architecture. We group these models into two categories: those for general image segmentation, and those for medical image segmentation.

3.3.1 General Image Segmentation

Noh *et al.* [41] introduced semantic segmentation based on deconvolution (a.k.a. transposed convolution). Their model, DeConvNet (Fig. 11), consists of two parts, an encoder using convolutional layers adopted from the VGG 16-layer network and a multilayer deconvolutional network that inputs the feature vector and generates a map of pixel-accurate class probabilities. The latter comprises deconvolution and unpooling layers, which identify pixel-wise class labels and predict segmentation masks.

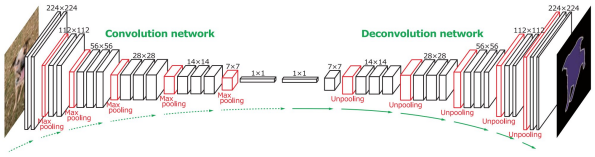


Fig. 11. Deconvolutional semantic segmentation. From [41].

Badrinarayanan *et al.* [25] proposed SegNet, a fully convolutional encoder-decoder architecture for image segmentation (Fig. 12). Similar to the deconvolution network, the core trainable segmentation engine of SegNet consists of an encoder network, which is topologically identical to the 13 convolutional layers of the VGG16 network, and a corresponding decoder network followed by a pixel-wise classification layer. The main novelty of SegNet is in the way the decoder upsamples its lower-resolution input feature map(s); specifically, using pooling indices computed in the max-pooling step of the corresponding encoder to perform nonlinear up-sampling.

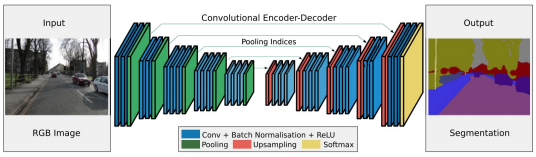


Fig. 12. The SegNet model. From [25].

A limitation of encoder-decoder based models is the loss of fine-grained image information, due to the loss of resolution through the encoding process. HRNet [42] (Fig. 13) addresses this shortcoming. Other than recovering high-resolution representations as is done in DeConvNet, SegNet, and other models, HRNet maintains high-resolution representations through the encoding process by connecting the high-to-low resolution convolution streams in parallel and repeatedly exchanging the information across resolutions. There are four stages: the 1st stage consists of high-resolution convolutions, while the 2nd/3rd/4th stage repeats 2-resolution/3-resolution/4-resolution blocks. Several recent semantic segmentation models use HRNet as a backbone.

Several other works adopt transposed convolutions, or encoder-decoders for image segmentation, such as Stacked Deconvolutional Network (SDN) [43], Linknet [44], W-Net [45], and locality-sensitive deconvolution networks for RGB-D segmentation [46].

3.3.2 Medical and Biomedical Image Segmentation

Several models inspired by FCNs and encoder-decoder networks were initially developed for medical/biomedical

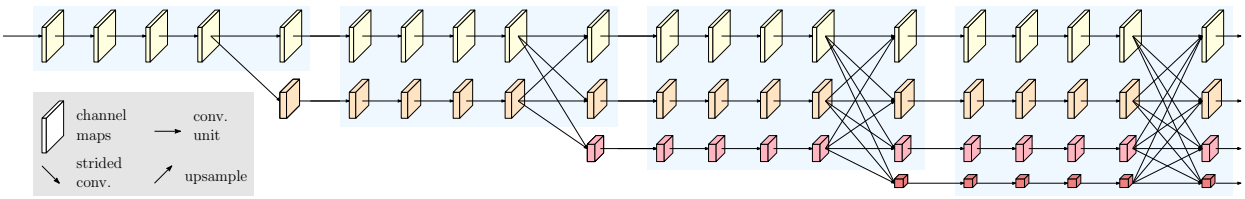


Fig. 13. The HRNet architecture. From [42].

image segmentation, but are now also being used outside the medical domain.

Ronneberger *et al.* [47] proposed the U-Net (Fig. 14) for efficiently segmenting biological microscopy images. The U-Net architecture comprises two parts, a contracting path to capture context, and a symmetric expanding path that enables precise localization. The U-Net training strategy relies on the use of data augmentation to learn effectively from very few annotated images. It was trained on 30 transmitted light microscopy images, and it won the ISBI cell tracking challenge 2015 by a large margin.

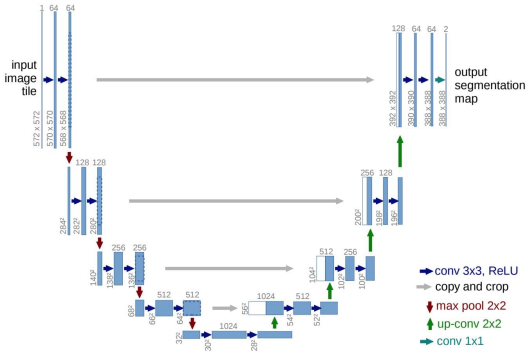


Fig. 14. The U-Net model. From [47].

Various extensions of U-Net have been developed for different kinds of images and problem domains; for example, Zhou *et al.* [48] developed a nested U-Net architecture, Zhang *et al.* [49] developed a road segmentation algorithm based on U-Net, and Cicek *et al.* [50] proposed a U-Net architecture for 3D images.

V-Net (Fig. 15), proposed by Milletari *et al.* [51] for 3D medical image segmentation, is another well known FCN-based model. The authors introduced a new loss function based on the Dice coefficient, enabling the model to deal with situations in which there is a strong imbalance between the number of voxels in the foreground and background. The network was trained end-to-end on MRI images of the prostate and learns to predict segmentation for the whole volume at once. Some of the other relevant works on medical image segmentation includes Progressive Dense V-Net *et al.* for automatic segmentation of pulmonary lobes from chest CT images, and the 3D-CNN encoder for lesion segmentation [52].

3.4 Multiscale and Pyramid Network Based Models

Multiscale analysis, a well established idea in image processing, has been deployed in various neural network architectures. One of the most prominent models of this

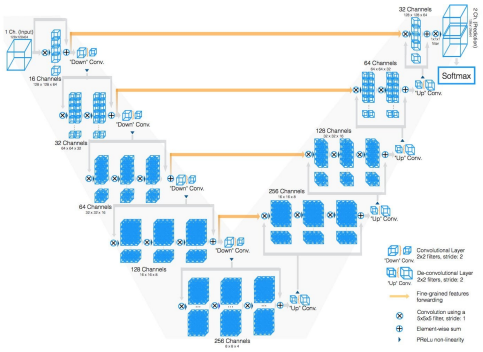


Fig. 15. The V-Net model for 3D image segmentation. From [51].

sort is the Feature Pyramid Network (FPN) proposed by Lin *et al.* [53], which was developed for object detection but was also applied to segmentation. The inherent multiscale, pyramidal hierarchy of deep CNNs was used to construct feature pyramids with marginal extra cost. To merge low and high resolution features, the FPN is composed of a bottom-up pathway, a top-down pathway and lateral connections. The concatenated feature maps are then processed by a 3×3 convolution to produce the output of each stage. Finally, each stage of the top-down pathway generates a prediction to detect an object. For image segmentation, the authors use two multilayer perceptrons (MLPs) to generate the masks.

Zhao *et al.* [54] developed the Pyramid Scene Parsing Network (PSPN), a multiscale network to better learn the global context representation of a scene (Fig. 16). Multiple patterns are extracted from the input image using a residual network (ResNet) as a feature extractor, with a dilated network. These feature maps are then fed into a pyramid pooling module to distinguish patterns of different scales. They are pooled at four different scales, each one corresponding to a pyramid level, and processed by a 1×1 convolutional layer to reduce their dimensions. The outputs of the pyramid levels are up-sampled and concatenated with the initial feature maps to capture both local and global context information. Finally, a convolutional layer is used to generate the pixel-wise predictions.

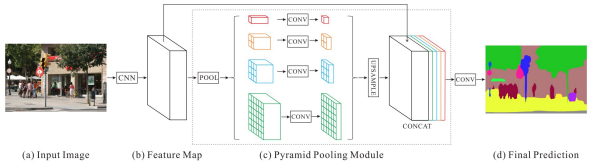


Fig. 16. The PSPN architecture. From [54].

Ghiasi and Fowlkes [55] developed a multiresolution

reconstruction architecture based on a Laplacian pyramid that uses skip connections from higher resolution feature maps and multiplicative gating to successively refine segment boundaries reconstructed from lower-resolution maps. They showed that while the apparent spatial resolution of convolutional feature maps is low, the high-dimensional feature representation contains significant sub-pixel localization information.

Other models use multiscale analysis for segmentation, among them Dynamic Multiscale Filters Network (DM-Net) [56], Context Contrasted Network and gated multiscale aggregation (CCN) [57], Adaptive Pyramid Context Network (APC-Net) [58], MultiScale Context Intertwining (MSCI) [59], and salient object segmentation [60].

3.5 R-CNN Based Models

The Regional CNN (R-CNN) and its extensions have proven successful in object detection applications. In particular, the Faster R-CNN [61] architecture (Fig. 17) uses a region proposal network (RPN) that proposes bounding box candidates. The RPN extracts a Region of Interest (RoI), and an RoIPool layer computes features from these proposals to infer the bounding box coordinates and class of the object. Some extensions of R-CNN have been used to address the instance segmentation problem; i.e., the task of simultaneously performing object detection and semantic segmentation.

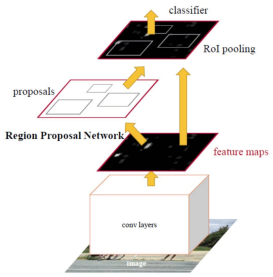


Fig. 17. Faster R-CNN architecture. Each image is processed by convolutional layers and its features are extracted, a sliding window is used in RPN for each location over the feature map, for each location, k ($k = 9$) anchor boxes are used (3 scales of 128, 256 and 512, and 3 aspect ratios of 1:1, 1:2, 2:1) to generate a region proposal; A cls layer outputs 2k scores whether there or not there is an object for k boxes; A reg layer outputs $4k$ for the coordinates (box center coordinates, width and height) of k boxes. From [61].

He *et al.* [62] proposed Mask R-CNN (Fig. 18), which outperformed previous benchmarks on many COCO object instance segmentation challenges (Fig. 19), efficiently detecting objects in an image while simultaneously generating a high-quality segmentation mask for each instance. Essentially, it is a Faster R-CNN with 3 output branches—the first computes the bounding box coordinates, the second computes the associated classes, and the third computes the binary mask to segment the object. The Mask R-CNN loss function combines the losses of the bounding box coordinates, the predicted class, and the segmentation mask, and trains all of them jointly.

The Path Aggregation Network (PANet) proposed by Liu *et al.* [63] is based on the Mask R-CNN and FPN models (Fig. 20). The feature extractor of the network uses an FPN backbone with a new augmented bottom-up pathway

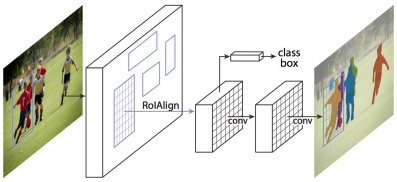


Fig. 18. Mask R-CNN architecture. From [62].

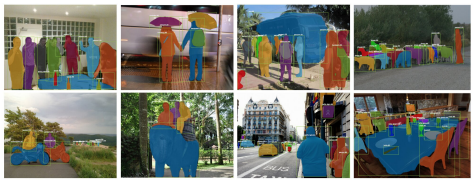


Fig. 19. Mask R-CNN instance segmentation results. From [62].

improving the propagation of lower-layer features. Each stage of this third pathway takes as input the feature maps of the previous stage and processes them with a 3×3 convolutional layer. A lateral connection adds the output to the same-stage feature maps of the top-down pathway and these feed the next stage.

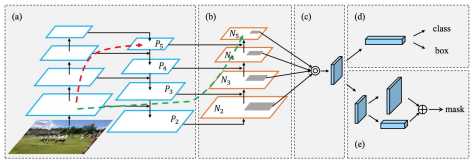


Fig. 20. The Path Aggregation Network. (a) FPN backbone. (b) Bottom-up path augmentation. (c) Adaptive feature pooling. (d) Box branch. (e) Fully-connected fusion. From [63].

Dai *et al.* [64] developed a multitask network for instance-aware semantic segmentation that consists of three networks for differentiating instances, estimating masks, and categorizing objects. These networks form a cascaded structure and are designed to share their convolutional features. Hu *et al.* [65] proposed a new partially-supervised training paradigm together with a novel weight transfer function, which enables training instance segmentation models on a large set of categories, all of which have box annotations, but only a small fraction of which have mask annotations.

Chen *et al.* [66] developed an instance segmentation model, MaskLab, by refining object detection with semantic and direction features based on Faster R-CNN. This model produces three outputs (Fig. 21), box detection, semantic segmentation logits for pixel-wise classification, and direction prediction logits for predicting each pixel's direction toward its instance center. Building on the Faster R-CNN object detector, the predicted boxes provide accurate localization of object instances. Within each region of interest, MaskLab performs foreground/background segmentation by combining semantic and direction prediction.

Tensormask, proposed by Chen *et al.* [67], is based on dense sliding window instance segmentation. The authors treat dense instance segmentation as a prediction task over 4D tensors and present a general framework that enables novel operators on 4D tensors. They demonstrate that the

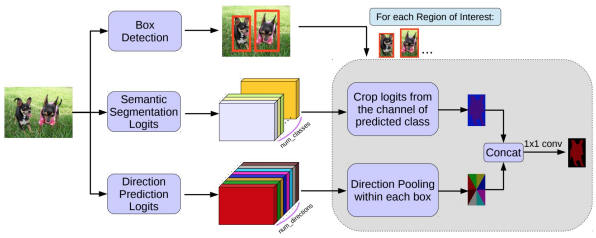


Fig. 21. The MaskLab model. From [66].

tensor approach yields large gains over baselines, with results comparable to Mask R-CNN.

Other instance segmentation models have been developed based on R-CNN, such as those developed for mask proposals, including R-FCN [68], DeepMask [69], PolarMask [70], boundary-aware instance segmentation [71], and CenterMask [72]. Another promising approach is to tackle the instance segmentation problem by learning grouping cues for bottom-up segmentation, such as deep watershed transform [73], real-time instance segmentation [74], and semantic instance segmentation via deep metric learning [75].

3.6 Dilated Convolutional Models

Dilated (a.k.a. "atrous") convolution introduces to convolutional layers another parameter, the dilation rate. For example, a 3×3 kernel (Fig. 22) with a dilation rate of 2 will have the same size receptive field as a 5×5 kernel while using only 9 parameters, thus enlarging the receptive field with no increase in computational cost.

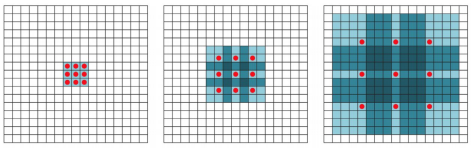


Fig. 22. Dilated convolution. A 3×3 kernel at different dilation rates.

Dilated convolutions have been popular in the field of real-time segmentation, and many recent publications report the use of this technique. Some of the most important include the DeepLab family [76], multiscale context aggregation [77], Dense Upsampling Convolution and Hybrid Dilated Convolution (DUC-HDC) [78], densely connected Atrous Spatial Pyramid Pooling (DenseASPP) [79], and the Efficient Network (ENet) [80].

DeepLabv1 [36] and DeepLabv2 [76], developed by Chen *et al.*, are among the most popular image segmentation models. The latter has three key features (Fig. 23). First is the use of dilated convolution to address the decreasing resolution in the network caused by max-pooling and striding. Second is Atrous Spatial Pyramid Pooling (ASPP), which probes an incoming convolutional feature layer with filters at multiple sampling rates, thus capturing objects as well as multiscale image context to robustly segment objects at multiple scales. Third is improved localization of object boundaries by combining methods from deep CNNs, such as fully convolutional VGG-16 or ResNet 101, and probabilistic graphical models, specifically fully-connected CRFs.

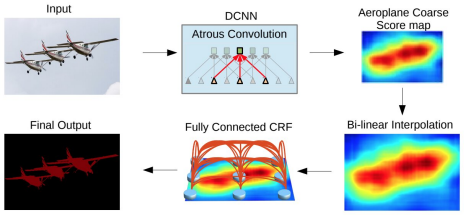


Fig. 23. The DeepLab model. From [76].

Subsequently, Chen *et al.* [13] proposed DeepLabv3, which combines cascaded and parallel modules of dilated convolutions. The parallel convolution modules are grouped in the ASPP. A 1×1 convolution and batch normalization are added in the ASPP. All the outputs are concatenated and processed by another 1×1 convolution to create the final output with logits for each pixel. Next, Chen *et al.* [81] released Deeplabv3+ (Fig. 24), which uses an encoder-decoder architecture including dilated separable convolution composed of a depthwise convolution (spatial convolution for each channel of the input) and pointwise convolution (1×1 convolution with the depthwise convolution as input). They used the DeepLabv3 framework as the encoder. The most relevant model has a modified Xception backbone with more layers, dilated depthwise separable convolutions instead of max pooling and batch normalization.

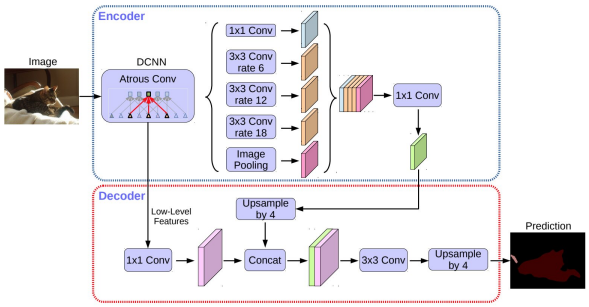


Fig. 24. The DeepLab-v3+ model. From [81].

3.7 RNN Based Models

While CNNs are a natural fit for computer vision problems, they are not the only possibility. RNNs are useful in modeling the short/long term dependencies among pixels to (potentially) improve the estimation of the segmentation map. Using RNNs, pixels may be linked together and processed sequentially to model global contexts and improve semantic segmentation. However the natural 2D structure of images poses a challenge.

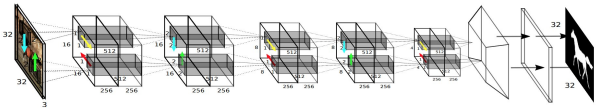


Fig. 25. The ReSeg model (without the pre-trained VGG-16 feature extractor). From [82].

Visin *et al.* [82] proposed an RNN-based model for semantic segmentation called ReSeg (Fig. 25). This model is mainly based on ReNet [83], which was developed for image classification. Each ReNet layer is composed of four RNNs that sweep the image horizontally and vertically in both directions, encoding patches/activations, and providing relevant global information. To perform image segmentation with the ReSeg model, ReNet layers are stacked atop pre-trained VGG-16 convolutional layers, which extract generic local features, and are then followed by up-sampling layers to recover the original image resolution in the final predictions.

Byeon *et al.* [84] performed per-pixel segmentation and classification of images of natural scenes using 2D LSTM networks, which learn textures and the complex spatial dependencies of labels in a single model that carries out classification, segmentation, and context integration.

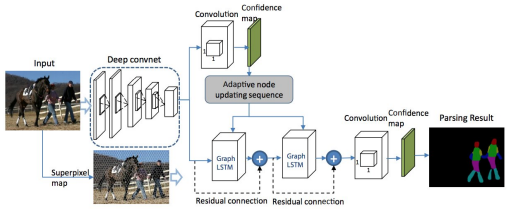


Fig. 26. The graph-LSTM model for semantic segmentation. From [85].

Liang *et al.* [85] proposed a semantic segmentation model based on a graph-LSTM network (Fig. 26) in which convolutional layers are augmented by graph-LSTM layers built on super-pixel maps, which provide a more global structural context. These layers generalize the LSTM for uniform, array-structured data (i.e., row, grid, or diagonal LSTMs) to nonuniform, graph-structured data, where arbitrary-shaped superpixels are semantically consistent nodes and the adjacency relations between superpixels correspond to edges, thus forming an undirected graph (Fig. 27).

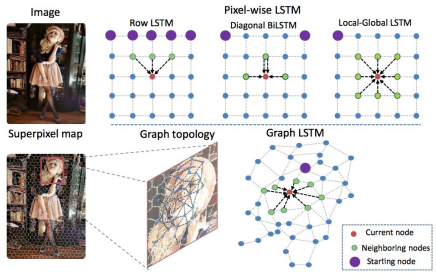


Fig. 27. Comparison of conventional RNN models and the graph-LSTM. From [85].

Xiang and Fox [86] proposed Data Associated Recurrent Neural Networks (DA-RNNs) for joint 3D scene mapping and semantic labeling. DA-RNNs use a new recurrent neural network architecture for semantic labeling on RGB-D videos. The output of the network is integrated with mapping techniques such as Kinect-Fusion in order to inject semantic information into the reconstructed 3D scene.

Hu *et al.* [87] developed a semantic segmentation algorithm that combines a CNN to encode the image and an LSTM to encode its linguistic description. To produce pixel-wise image segmentations from language inputs, they propose an end-to-end trainable recurrent and convolutional model that jointly learns to process visual and linguistic information (Fig. 28). This differs from traditional semantic segmentation over a predefined set of semantic classes; i.e., the phrase "two men sitting on the right bench" requires segmenting only the two people on the right bench and no others sitting on another bench or standing. Fig. 29 shows an example segmentation result by the model.

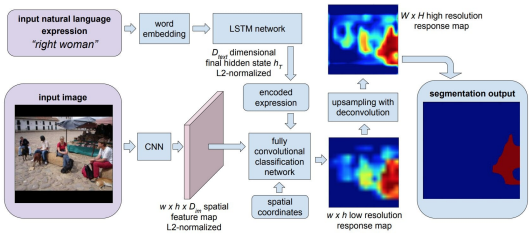


Fig. 28. The CNN+LSTM architecture for semantic segmentation from natural language expressions. From [87].

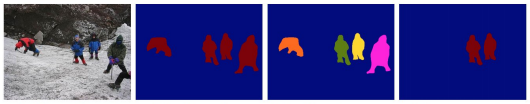


Fig. 29. CNN+LSTM segmentation masks generated for the query "people in blue coat". From [87].

A drawback of RNN-based models is that they will generally be slower than their CNN counterparts as their sequential nature is not amenable to parallelization.

3.8 Attention-Based Models

Attention mechanisms have been persistently explored in computer vision over the years, and it is not surprising to find publications that apply them to semantic segmentation.

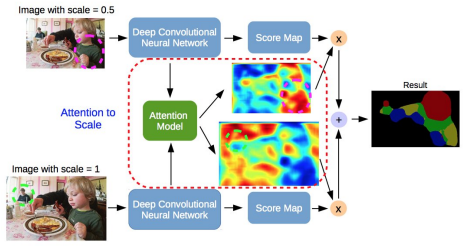


Fig. 30. Attention-based semantic segmentation model. From [88].

Chen *et al.* [88] proposed an attention mechanism that learns to softly weight multiscale features at each pixel location. They adapt a powerful semantic segmentation model and jointly train it with multiscale images and the attention model In Fig. 30, the model assigns large weights to the person (green dashed circle) in the background for features from scale 1.0 as well as on the large child (magenta dashed circle) for features from scale 0.5. The attention mechanism enables the model to assess the importance of features at different positions and scales, and it outperforms average and max pooling.

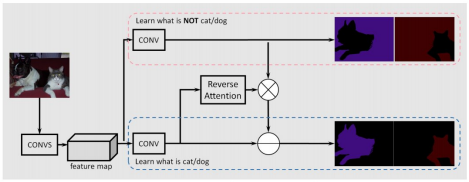


Fig. 31. The RAN architecture. From [89].

Unlike approaches in which convolutional classifiers are trained to learn the representative semantic features of labeled objects, Huang *et al.* [89] proposed a Reverse Attention Network (RAN) architecture (Fig. 31) for semantic segmentation that also applies reverse attention mechanisms, thereby training the model to capture the opposite concept—features that are not associated with a target class. The RAN network performs the direct and reverse-attention learning processes simultaneously.

Li *et al.* [90] developed a Pyramid Attention Network for semantic segmentation, which exploits global contextual information for semantic segmentation. Eschewing complicated dilated convolutions and decoder networks, they combined attention mechanisms and spatial pyramids to extract precise dense features for pixel labeling. Fu *et al.* [91] proposed a dual attention network for scene segmentation that can capture rich contextual dependencies based on the self-attention mechanism. Specifically, they append two types of attention modules on top of a dilated FCN that models the semantic inter-dependencies in spatial and channel dimensions, respectively. The position attention module selectively aggregates the features at each position via weighted sums.

Other applications of attention mechanisms to semantic segmentation include OCNet [92], which employs an object context pooling inspired by self-attention mechanism, ResNeSt: Split-Attention Networks [93], Height-driven Attention Networks [94], Expectation-Maximization Attention (EMANet) [95], Criss-Cross Attention Network (CCNet) [96], end-to-end instance segmentation with recurrent attention [97], a point-wise spatial attention network for scene parsing [98], and Discriminative Feature Network (DFN) [99].

3.9 Generative Models and Adversarial Training

GANs have been applied to a wide range of tasks in computer vision, not excluding image segmentation.

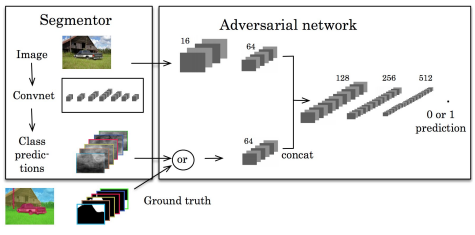


Fig. 32. The GAN for semantic segmentation. From [100].

Luc *et al.* [100] proposed an adversarial training approach for semantic segmentation in which they trained a convolutional semantic segmentation network (Fig. 32), along with an adversarial network that discriminates between ground-truth segmentation maps and those generated by the segmentation network. They showed that the adversarial training approach yields improved accuracy on the Stanford Background and PASCAL VOC 2012 datasets.

Souly *et al.* [101] proposed semi-weakly supervised semantic segmentation using GANs. Their model consists of a generator network providing extra training examples to a multiclass classifier, acting as discriminator in the GAN framework, that assigns sample a label from the possible label classes or marks it as a fake sample (extra class).

Hung *et al.* [102] developed a framework for semi-supervised semantic segmentation using an adversarial network. They designed an FCN discriminator to differentiate the predicted probability maps from the ground truth segmentation distribution, considering the spatial resolution. The loss function of this model has three terms: cross-entropy loss on the segmentation ground truth, adversarial loss of the discriminator network, and semi-supervised loss based on the confidence map output of the discriminator.

Xue *et al.* [103] proposed an adversarial network with multiscale L1 Loss for medical image segmentation. They used an FCN as the segmentor to generate segmentation label maps, and proposed a novel adversarial critic network with a multi-scale L1 loss function to force the critic and segmentor to learn both global and local features that capture long and short range spatial relationships between pixels.

Other approaches based on adversarial training include cell image segmentation using GANs [104], and segmentation and generation of the invisible parts of objects [105].

3.10 CNN Models With Active Contour Models

The exploration of synergies between FCNs and Active Contour Models (ACMs) [8] has recently attracted research interest.

One approach is to formulate new loss functions that are inspired by ACM principles. For example, inspired by the global energy formulation of [106], Chen *et al.* [107] proposed a supervised loss layer that incorporated area and size information of the predicted masks during training of an FCN and tackled the problem of ventricle segmentation in cardiac MRI. Similarly, Gur *et al.* [108] presented an unsupervised loss function based on morphological active contours without edges [109] for microvascular image segmentation.

A different approach initially sought to utilize the ACM merely as a post-processor of the output of an FCN and several efforts attempted modest co-learning by pre-training the FCN. One example of an ACM post-processor for the task of semantic segmentation of natural images is the work by Le *et al.* [110] in which level-set ACMs are implemented as RNNs. Deep Active Contours by Rupprecht *et al.* [111], is another example. For medical image segmentation, Hatamizadeh *et al.* [112] proposed an integrated Deep Active Lesion Segmentation (DALS) model that trains the FCN backbone to predict the parameter functions of a novel, locally-parameterized level-set energy functional. In another relevant effort, Marcos *et al.* [113] proposed Deep Structured Active Contours (DSAC), which combines ACMs and pre-trained FCNs in a structured prediction framework for building instance segmentation (albeit with manual initialization) in aerial images. For the same application,

Cheng *et al.* [114] proposed the Deep Active Ray Network (DarNet), which is similar to DSAC, but with a different explicit ACM formulation based on polar coordinates to prevent contour self-intersection.

A truly end-to-end backpropagation trainable, fully-integrated FCN-ACM combination was recently introduced by Hatamizadeh *et al.* [115], dubbed Trainable Deep Active Contours (TDAC). Going beyond [112], they implemented the locally-parameterized level-set ACM in the form of additional convolutional layers following the layers of the backbone FCN, exploiting Tensorflow's automatic differentiation mechanism to backpropagate training error gradients throughout the entire DCAC framework. The fully-automated model requires no intervention either during training or segmentation, can naturally segment multiple instances of objects of interest, and deal with arbitrary object shape including sharp corners.

3.11 Other Models

Other popular DL architectures for image segmentation include the following:

Context Encoding Network (EncNet) [116] uses a basic feature extractor and feeds the feature maps into a context encoding module. RefineNet [117] is a multipath refinement network that explicitly exploits all the information available along the down-sampling process to enable high-resolution prediction using long-range residual connections. Seednet [118] introduced an automatic seed generation technique with deep reinforcement learning that learns to solve the interactive segmentation problem. Object-Contextual Representations (OCR) [42] learns object regions and the relation between each pixel and each object region, augmenting the representation pixels with the object-contextual representation. Additional models and methods include BoxSup [119], Graph Convolutional Networks (GCN) [120], Wide ResNet [121], Exfuse [122] (enhancing low-level and high-level features fusion), Feedforward-Net [123], saliency-aware models for geodesic video segmentation [124], Dual Image Segmentation (DIS) [125], FoveaNet [126] (perspective-aware scene parsing), Ladder DenseNet [127], Bilateral Segmentation Network (BiSeNet) [128], Semantic Prediction Guidance for Scene Parsing (SPGNet) [129], gated shape CNNs [130], Adaptive Context Network (AC-Net) [131], Dynamic-Structured Semantic Propagation Network (DSSPN) [132], Symbolic Graph Reasoning (SGR) [133], CascadeNet [134], Scale-Adaptive Convolutions (SAC) [135], Unified Perceptual parsing Network (UperNet) [136], segmentation by re-training and self-training [137], densely connected neural architecture search [138], hierarchical multiscale attention [139], Efficient RGB-D Semantic Segmentation (ESA-Net) [140], Iterative Pyramid Contexts [141], and Learning Dynamic Routing for Semantic Segmentation [142].

Panoptic segmentation [143] is growing in popularity. Efforts in this direction include Panoptic Feature Pyramid Network (PFPN) [144], attention-guided network for panoptic segmentation [145], seamless scene segmentation [146], panoptic Deeplab [147], unified panoptic segmentation network [148], and efficient panoptic segmentation [149].

Fig. 33 provides a timeline of some of the most representative DL image segmentation models since 2014.

4 Datasets

In this section we survey the image datasets most commonly used to train and test DL image segmentation models, grouping them into 3 categories—2D (pixel) images, 2.5D RGB-D (color+depth) images, and 3D (voxel) images—and provide details about the characteristics of each dataset.

Data augmentation is often used to increase the number of labeled samples, especially for small datasets such as those in the medical imaging domain, thus improving the performance of DL segmentation models. A set of transformations is applied either in the data space, or feature space, or both (i.e., both the image and the segmentation map). Typical transformations include translation, reflection, rotation, warping, scaling, color space shifting, cropping, and projections onto principal components. Data augmentation can also benefit by yielding faster convergence, decreasing the chance of over-fitting, and enhancing generalization. For some small datasets, data augmentation has been shown to boost model performance by more than 20%.

4.1 2D Image Datasets

The bulk of image segmentation research has focused on 2D images; therefore, many 2D image segmentation datasets are available. The following are some of the most popular:

PASCAL Visual Object Classes (VOC) [150] is a highly popular dataset in computer vision, with annotated images available for 5 tasks—classification, segmentation, detection, action recognition, and person layout. For the segmentation task, there are 21 labeled object classes and pixels are labeled as background if they do not belong to any of these classes. The dataset is divided into two sets, training and validation, with 1,464 and 1,449 images, respectively, and a private test set for the actual challenge. Fig. 34 shows an example image and its pixel-wise label.

PASCAL Context [152] is an extension of the PASCAL VOC 2010 detection challenge. It includes pixel-wise labels for all the training images. It contains more than 400 classes (including the original 20 classes plus backgrounds from PASCAL VOC segmentation), in three categories (objects, stuff, and hybrids). Many of the object categories of this dataset are too sparse and; therefore, a subset of 59 classes is usually selected for use.

Microsoft Common Objects in Context (MS COCO) [153] is a large-scale object detection, segmentation, and captioning dataset. COCO includes images of complex everyday scenes, containing common objects in their natural contexts. This dataset contains photos of 91 object types, with a total of 2.5 million labeled instances in 328K images. Fig. 35 compares MS-COCO labels with those of previous datasets for a sample image.

Cityscapes [154] is a large database with a focus on semantic understanding of urban street scenes. It contains a diverse set of stereo video sequences recorded in street scenes from 50 cities, with high quality pixel-level annotation of 5K frames, in addition to a set of 20K weakly annotated frames. It includes semantic and dense pixel annotations of 30 classes, grouped into 8 categories—flat surfaces, humans, vehicles, constructions, objects, nature, sky, and void. Fig. 36 shows sample segmentation maps from this dataset.

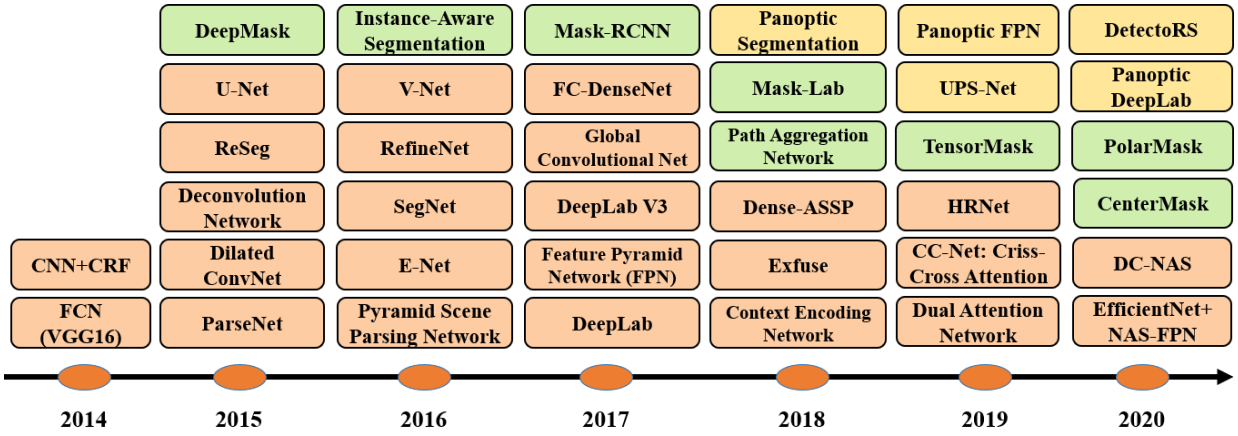


Fig. 33. Timeline of representative DL-based image segmentation algorithms. Orange, green, and yellow blocks indicate semantic, instance, and panoptic segmentation algorithms, respectively.

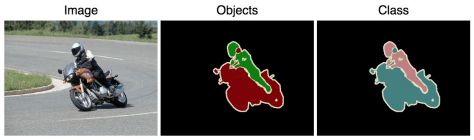


Fig. 34. An example image from the PASCAL VOC dataset. From [151].

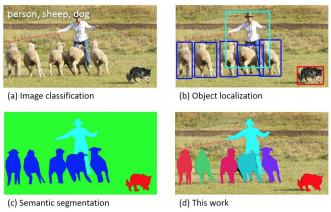


Fig. 35. A sample image and segmentation map in COCO. From [153].

ADE20K/MIT Scene Parsing (SceneParse150) offers a training and evaluation platform for scene parsing algorithms. The data for this benchmark comes from the ADE20K dataset [134], which contains more than 20K scene-centric images exhaustively annotated with objects and object parts. The benchmark is divided into 20K images for training, 2K images for validation, and another batch of images for testing. There are 150 semantic categories in this dataset.

SiftFlow [155] includes 2,688 annotated images, from a subset of the LabelMe database, of 8 different outdoor scenes, among them streets, mountains, fields, beaches, and buildings, and in one of 33 semantic classes.

Stanford Background [156] comprises outdoor images of scenes from existing datasets, such as LabelMe, MSRC, and PASCAL VOC. It includes 715 images with at least one foreground object. The dataset is pixel-wise annotated, and

Fig. 36. Segmentation maps from the Cityscapes dataset. From [154].

can be used for semantic scene understanding.

Berkeley Segmentation Dataset (BSD) [157] contains 12,000 hand-labeled segmentations of 1,000 Corel dataset images from 30 human subjects. It aims to provide an empirical basis for research on image segmentation and boundary detection. Half of the segmentations were obtained from presenting the subject a color image and the other half from presenting a grayscale image. The public benchmark based on this data consists of all of the grayscale and color segmentations for 300 images. The images are divided into a training set of 200 images and a test set of 100 images.

Youtube-Objects [158] contains videos collected from YouTube, which include objects from ten PASCAL VOC classes (aeroplane, bird, boat, car, cat, cow, dog, horse, motorbike, and train). The original dataset did not contain pixel-wise annotations (as it was originally developed for object detection, with weak annotations). However, Jain *et al.* [159] manually annotated a subset of 126 sequences, and then extracted a subset of frames to further generate semantic labels. In total, there are about 10,167 annotated 480x360 pixel frames available in this dataset.

CamVid: is another scene understanding database (with a focus on road/driving scenes) which was originally captured as five video sequences via camera mounted on the dashboard of a car. A total of 701 frames were provided by sampling from the sequences. These frames were manually annotated into 32 classes.

KITTI [160] is one of the most popular datasets for autonomous driving, containing videos of traffic scenarios, recorded with a variety of sensor modalities (including high-resolution RGB, grayscale stereo cameras, and a 3D laser scanners). The original dataset does not contain ground truth for semantic segmentation, but researchers have manually annotated parts of the dataset; e.g., Alvarez *et al.* [161] generated ground truth for 323 images from the road detection challenge with 3 classes—road, vertical, and sky.

Other datasets for image segmentation purposes include **Semantic Boundaries Dataset (SBD)** [162], **PASCAL Part** [163], **SYNTHIA** [164], and **Adobe's Portrait Segmentation** [165].

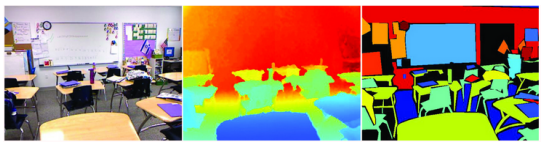


Fig. 37. A sample from the NYU V2 dataset. From left: RGB image, pre-processed depth image, class labels map. From [166].

4.2 2.5D Datasets

With the availability of affordable range scanners, RGB-D images have became increasingly widespread. The following RGB-D datasets are among the most popular:

NYU-Depth V2 [166] consists of video sequences from a variety of indoor scenes, recorded by the RGB and depth cameras of the Microsoft Kinect. It includes 1,449 densely labeled RGB and depth image pairs of more than 450 scenes taken from 3 cities. Each object is labeled with a class and instance number (e.g., cup1, cup2, cup3, etc.). It also contains 407,024 unlabeled frames. Fig. 37 shows an RGB-D sample and its label map.

SUN-3D [167] is a large RGB-D video dataset that contains 415 sequences captured from 254 different spaces in 41 different buildings; 8 sequences are annotated and more will be annotated in the future. Each annotated frame provides the semantic segmentation of the objects in the scene as well as information about the camera pose.

SUN RGB-D [168] provides an RGB-D benchmark for advancing the state-of-the-art of all major scene understanding tasks. It is captured by four different sensors and contains 10,000 RGB-D images at a scale similar to PASCAL VOC.

ScanNet [169] is an RGB-D video dataset containing 2.5 million views in more than 1,500 scans, annotated with 3D camera poses, surface reconstructions, and instance-level semantic segmentations. To collect these data, an easy-to-use and scalable RGB-D capture system was designed that includes automated surface reconstruction, and the semantic annotation was crowd-sourced. Using this data helped achieve state-of-the-art performance on several 3D scene understanding tasks, including 3D object classification, semantic voxel labeling, and CAD model retrieval.

Stanford 2D-3D [170] provides a variety of mutually registered 2D, 2.5D, and 3D modalities, with instance-level semantic and geometric annotations, acquired from 6 indoor areas. It contains over 70,000 RGB images, along with the corresponding depths, surface normals, semantic annotations, as well as global XYZ images, camera information, and registered raw and semantically annotated 3D meshes and point clouds.

Another popular 2.5D datasets is **UW RGB-D Object Dataset** [171], which contains 300 common household objects recorded using a Kinect-style sensor.

5 DL Segmentation Model Performance

In this section, we summarize the metrics commonly used in evaluating the performance of segmentation models and report the performance of DL-based segmentation models on benchmark datasets.

5.1 Metrics for Image Segmentation Models

Ideally, an image segmentation model should be evaluated in multiple respects, such as quantitative accuracy, visual quality, speed (inference time), and storage requirements (memory footprint). However, most researchers to date have focused on metrics for quantifying model accuracy. The following metrics are most popular:

Pixel accuracy is the ratio of properly classified pixels divided by the total number of pixels. For $K+1$ classes (K foreground classes and the background) pixel accuracy is defined as

$$\text{PA} = \frac{\sum_{i=0}^{K} p_{ii}}{\sum_{i=0}^{K}\sum_{j=0}^{K} p_{ij}}, \tag{1}$$

where p_{ij} is the number of pixels of class i predicted as belonging to class j.

Mean Pixel Accuracy (MPA) is an extension of PA, in which the ratio of correct pixels is computed in a per-class manner and then averaged over the total number of classes:

$$\text{MPA} = \frac{1}{K+1}\sum_{i=0}^{K}\frac{p_{ii}}{\sum_{j=0}^{K} p_{ij}}. \tag{2}$$

Intersection over Union (IoU), or the **Jaccard Index**, is defined as the area of intersection between the predicted segmentation map A and the ground truth map B, divided by the area of the union between the two maps, and ranges between 0 and 1:

$$\text{IoU} = J(A,B) = \frac{|A \cap B|}{|A \cup B|}. \tag{3}$$

Mean-IoU is defined as the average IoU over all classes.

Precision / Recall / F1 score can be defined for each class, as well as at the aggregate level, as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}; \qquad \text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}, \tag{4}$$

where TP refers to the true positive fraction, FP refers to the false positive fraction, and FN refers to the false negative fraction. Usually one is interested in a combined version of precision and recall rates; the F1 score is defined as the harmonic mean of precision and recall:

$$\text{F1} = \frac{2\,\text{Precision}\,\text{Recall}}{\text{Precision}+\text{Recall}}. \tag{5}$$

Dice coefficient, commonly used in medical image analysis, can be defined as twice the overlap area of the predicted and ground-truth maps divided by the total number of pixels.

$$\text{Dice} = \frac{2|A \cap B|}{|A|+|B|}. \tag{6}$$

It is very similar to the IoU (3) and when applied to binary maps, with foreground as the positive class, the Dice coefficient is identical to the F1 score (7):

$$\text{Dice} = \frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}} = \text{F1}. \tag{7}$$

TABLE 1
Accuracies of segmentation models on the PASCAL VOC test set

Method	Backbone	mIoU
FCN [30]	VGG-16	62.2
CRF-RNN [38]	-	72.0
CRF-RNN* [38]	-	74.7
BoxSup* [119]	-	75.1
Piecewise* [39]	-	78.0
DPN* [40]	-	77.5
DeepLab-CRF [76]	ResNet-101	79.7
GCN* [120]	ResNet-152	82.2
Dynamic Routing [142]	-	84.0
RefineNet [117]	ResNet-152	84.2
Wide ResNet [121]	WideResNet-38	84.9
PSPNet [54]	ResNet-101	85.4
DeeplabV3 [13]	ResNet-101	85.7
PSANet [98]	ResNet-101	85.7
EncNet [116]	ResNet-101	85.9
DFN* [99]	ResNet-101	86.2
Exfuse [122]	ResNet-101	86.2
SDN* [43]	DenseNet-161	86.6
DIS [125]	ResNet-101	86.8
APC-Net* [58]	ResNet-101	87.1
EMANet [95]	ResNet-101	87.7
DeeplabV3+ [81]	Xception-71	87.8
Exfuse [122]	ResNeXt-131	87.9
MSCI [59]	ResNet-152	88.0
EMANet [95]	ResNet-152	88.2
DeeplabV3+* [81]	Xception-71	89.0
EfficientNet+NAS-FPN [137]	-	90.5

* Models pre-trained on other datasets (MS-COCO, ImageNet, etc.).

TABLE 2
Accuracies of segmentation models on the Cityscapes dataset

Method	Backbone	mIoU
SegNet [25]	-	57.0
FCN-8s [30]	-	65.3
DPN [40]	-	66.8
Dilation10 [77]	-	67.1
DeeplabV2 [76]	ResNet-101	70.4
RefineNet [117]	ResNet-101	73.6
FoveaNet [126]	ResNet-101	74.1
Ladder DenseNet [127]	Ladder DenseNet-169	73.7
GCN [120]	ResNet-101	76.9
DUC-HDC [78]	ResNet-101	77.6
Wide ResNet [121]	WideResNet-38	78.4
PSPNet [54]	ResNet-101	85.4
BiSeNet [128]	ResNet-101	78.9
DFN [99]	ResNet-101	79.3
PSANet [98]	ResNet-101	80.1
DenseASPP [79]	DenseNet-161	80.6
Dynamic Routing [142]	-	80.7
SPGNet [129]	2xResNet-50	81.1
DANet [91]	ResNet-101	81.5
CCNet [96]	ResNet-101	81.4
DeeplabV3 [13]	ResNet-101	81.3
IPC [141]	ResNet-101	81.8
AC-Net [131]	ResNet-101	82.3
OCR [42]	ResNet-101	82.4
ResNeSt200 [93]	ResNeSt-200	82.7
GS-CNN [130]	WideResNet	82.8
HA-Net [94]	ResNext-101	83.2
HRNetV2+OCR [42]	HRNetV2-W48	83.7
Hierarchical MSA [139]	HRNet-OCR	85.1

5.2 Quantitative Performance of DL-Based Models

In this section we tabulate the performance of several of the previously discussed algorithms on popular segmentation benchmarks. Although most publications report model performance on standard datasets and use standard metrics, some of them fail to do so, making across-the-board comparisons difficult. Furthermore, only a few publications provide additional information, such as execution time and memory footprint, in a reproducible way, which is important to industrial applications (such as drones, self-driving cars, robotics, etc.) that may run on embedded systems with limited computational power and storage, thus requiring light-weight models.

The following tables summarize the performances of several of the prominent DL-based segmentation models on different datasets:

Table 1 focuses on the PASCAL VOC test set. Clearly, there has been much improvement in the accuracy of the models since the introduction of the first DL-based image segmentation model, the FCN.

Table 2 focuses on the Cityscape test dataset. The latest models feature about 23% relative gain over the pioneering FCN model on this dataset.

Table 3 focuses on the MS COCO stuff test set. This dataset is more challenging than PASCAL VOC, and Cityescapes, as the highest mIoU is approximately 40%.

Table 4 focuses on the ADE20k validation set. This dataset is also more challenging than the PASCAL VOC and Cityescapes datasets.

Table 5 provides the performance of prominent instance segmentation algorithms on COCO test-dev 2017 dataset, in terms of average precision, and their speed.

TABLE 3
Accuracies of segmentation models on the MS COCO stuff dataset

Method	Backbone	mIoU
RefineNet [117]	ResNet-101	33.6
CCN [57]	Ladder DenseNet-101	35.7
DANet [91]	ResNet-50	37.9
DSSPN [132]	ResNet-101	37.3
EMA-Net [95]	ResNet-50	37.5
SGR [133]	ResNet-101	39.1
OCR [42]	ResNet-101	39.5
DANet [91]	ResNet-101	39.7
EMA-Net [95]	ResNet-50	39.9
AC-Net [131]	ResNet-101	40.1
OCR [42]	HRNetV2-W48	40.5

Table 6 provides the performance of prominent panoptic segmentation algorithms on MS-COCO val dataset, in terms of panoptic quality [143].

Finally, Table 7 summarizes the performance of several prominent models for RGB-D segmentation on the NYUD-v2 and SUN-RGBD datasets.

In summary, we have witnessed been significant improvement in the performance of deep segmentation models over the past 5–6 years, with a relative improvement of 25%-42% in mIoU on different datasets. However, some publications suffer from lack of reproducibility for multiple reasons—they report performance on non-standard benchmarks/databases, or only on arbitrary subsets of the test set from a popular benchmark, or they do not adequately describe the experimental setup and sometimes evaluate model performance only on a subset of object classes. Most importantly, many publications do not provide the source-

TABLE 4
Accuracies of segmentation models on the ADE20k validation dataset

Method	Backbone	mIoU
FCN [30]	-	29.39
DilatedNet [77]	-	32.31
CascadeNet [134]	-	34.90
RefineNet [117]	ResNet-152	40.7
PSPNet [54]	ResNet-101	43.29
PSPNet [54]	ResNet-269	44.94
EncNet [116]	ResNet-101	44.64
SAC [135]	ResNet-101	44.3
PSANet [98]	ResNet-101	43.70
UperNet [136]	ResNet-101	42.66
DSSPN [132]	ResNet-101	43.68
DM-Net [56]	ResNet-101	45.50
AC-Net [131]	ResNet-101	45.90
ResNeSt-101 [93]	ResNeSt-101	46.91
ResNeSt-200 [93]	ResNeSt-200	48.36

TABLE 5
Instance segmentation model performance on COCO test-dev 2017

Method	Backbone	FPS	AP
YOLACT-550 [74]	R-101-FPN	33.5	29.8
YOLACT-700 [74]	R-101-FPN	23.8	31.2
RetinaMask [172]	R-101-FPN	10.2	34.7
TensorMask [67]	R-101-FPN	2.6	37.1
SharpMask [173]	R-101-FPN	8.0	37.4
Mask-RCNN [62]	R-101-FPN	10.6	37.9
CenterMask [72]	R-101-FPN	13.2	38.3

code for their model implementations. Fortunately, with the increasing popularity of deep learning models, the trend has been positive and many research groups are moving toward reproducible frameworks and open-sourcing their implementations.

6 Challenges and Opportunities

Without a doubt, image segmentation has benefited greatly from deep learning, but several challenges lie ahead. We will next discuss some of the promising research directions that we believe will help in further advancing image segmentation algorithms.

TABLE 6
Panoptic segmentation model performance on MS-COCO val

Method	Backbone	PQ
Panoptic FPN [144]	ResNet-50	39.0
Panoptic FPN [144]	ResNet-101	40.3
AU-Net [145]	ResNet-50	39.6
Panoptic-DeepLab [147]	Xception-71	39.7
OANet [174]	ResNet-50	39.0
OANet [174]	ResNet-101	40.7
AdaptIS [175]	ResNet-50	35.9
AdaptIS [175]	ResNet-101	37.0
UPSNet* [148]	ResNet-50	42.5
OCFusion* [176]	ResNet-50	41.3
OCFusion* [176]	ResNet-101	43.0
OCFusion* [176]	ResNeXt-101	45.7

* Use of deformable convolution.

TABLE 7
Segmentation model performance on the NYUD-v2 and SUN-RGBD

	NYUD-v2		SUN-RGBD	
Method	m-Acc	m-IoU	m-Acc	m-IoU
Mutex [177]	-	31.5	-	-
MS-CNN [178]	45.1	34.1	-	-
FCN [30]	46.1	34.0	-	-
Joint-Seg [179]	52.3	39.2	-	-
SegNet [25]	-	-	44.76	31.84
Structured Net [39]	53.6	40.6	53.4	42.3
B-SegNet [180]	-	-	45.9	30.7
3D-GNN [181]	55.7	43.1	57.0	45.9
LSD-Net [46]	60.7	45.9	58.0	-
RefineNet [117]	58.9	46.5	58.5	45.9
D-aware CNN [182]	61.1	48.4	53.5	42.0
MTI-Net [183]	62.9	49	-	-
RDFNet [184]	62.8	50.1	60.1	47.7
ESANet-R34-NBt1D [140]	-	50.3	-	48.17
G-Aware Net [185]	68.7	59.6	74.9	54.5

6.1 More Challenging Datasets

Several large-scale image datasets have been created for semantic segmentation and instance segmentation. However, there remains a need for more challenging datasets, as well as datasets of different kinds of images. For still images, datasets with a large number of objects and overlapping objects would be very valuable. This can enable the training of models that handle dense object scenarios better, as well as large overlaps among objects as is common in real-world scenarios. With the rising popularity of 3D image segmentation, especially in medical image analysis, there is also a strong need for large-scale annotated 3D image datasets, which are more difficult to create than their lower dimensional counterparts.

6.2 Combining DL and Earlier Segmentation Models

There is now broad agreement that the performance of DL-based segmentation algorithms is plateauing, especially in certain application domains such as medical image analysis. To advance to the next level of performance, we must further explore the combination of CNN-based image segmentation models with prominent "classical" model-based image segmentation methods. The integration of CNNs with graphical models has been studied, but their integration with active contours, graph cuts, and other segmentation models is fairly recent and deserves further work.

6.3 Interpretable Deep Models

While DL-based models have achieved promising performance on challenging benchmarks, there remain open questions about these models. For example, what exactly are deep models learning? How should we interpret the features learned by these models? What is a minimal neural architecture that can achieve a certain segmentation accuracy on a given dataset? Although some techniques are available to visualize the learned convolutional kernels of these models, a comprehensive study of the underlying behavior/dynamics of these models is lacking. A better understanding of the theoretical aspects of these models can enable the development of better models curated toward various segmentation scenarios.

6.4 Weakly-Supervised and Unsupervised Learning

Weakly-supervised (a.k.a. few shot) learning [186] and unsupervised learning [187] are becoming very active research areas. These techniques promise to be specially valuable for image segmentation, as collecting pixel-accurately labeled training images is problematic in many application domains, particularly so in medical image analysis. The transfer learning approach is to train a generic image segmentation model on a large set of labeled samples (perhaps from a public benchmark) and then fine-tune that model on a few samples from some specific target application. Self-supervised learning is another promising direction that is attracting much attraction in various fields. With the help of self-supervised learning, many details in images can be captured in order to train segmentation models with far fewer training samples. Models based on reinforcement learning could also be another potential future direction, as they have scarcely received attention for image segmentation. For example, MOREL [188] introduced a deep reinforcement learning approach for moving object segmentation in videos.

6.5 Real-time Models for Various Applications

In many applications, accuracy is the most important factor; however, there are applications in which it is also critical to have segmentation models that can run in near real-time, or at common camera frame rates (at least 25 frames per second). This is useful for computer vision systems that are, for example, deployed in autonomous vehicles. Most of the current models are far from this frame-rate; e.g., FCN-8 takes roughly 100 ms to process a low-resolution image. Models based on dilated convolution help to increase the speed of segmentation models to some extent, but there is still plenty of room for improvement.

6.6 Memory Efficient Models

Many modern segmentation models require a significant amount of memory even during the inference stage. So far, much effort has been directed towards improving the accuracy of such models, but in order to fit them into specific devices, such as mobile phones, the networks must be simplified. This can be done either by using simpler models, or by using model compression techniques, or even by training a complex model and using knowledge distillation techniques to compress it into a smaller, memory efficient network that mimics the complex model.

6.7 Applications

DL-based segmentation methods have been successfully applied to satellite images in remote sensing [189], such as to support urban planning [190] and precision agriculture [191]. Images collected by airborne platforms [192] and drones [193] have also been segmented using DL-based segmentation methods in order to address important environmental problems including ones related to climate change. The main challenges of the remote sensing domain stem from the typically formidable size of the imagery (often collected by imaging spectrometers with hundreds or even thousands of spectral bands) and the limited ground-truth information necessary to evaluate the accuracy of the segmentation algorithms. Similarly, DL-based segmentation techniques in the evaluation of construction materials [194] face challenges related to the massive volume of the related image data and the limited reference information for validation purposes. Last but not least, an important application field for DL-based segmentation has been biomedical imaging [195]. Here, an opportunity is to design standardized image databases useful in evaluating new infectious diseases and tracking pandemics [196].

7 Conclusions

We have surveyed image segmentation algorithms based on deep learning models, which have achieved impressive performance in various image segmentation tasks and benchmarks, grouped into architectural categories such as: CNN and FCN, RNN, R-CNN, dilated CNN, attention-based models, generative and adversarial models, among others. We have summarized the quantitative performance of these models on some popular benchmarks, such as the PASCAL VOC, MS COCO, Cityscapes, and ADE20k datasets. Finally, we discussed some of the open challenges and promising research directions for deep-learning-based image segmentation in the coming years.

Acknowledgments

We thank Tsung-Yi Lin of Google Brain as well as Jingdong Wang and Yuhui Yuan of Microsoft Research Asia for providing helpful comments that improved the manuscript.

References

[1] A. Rosenfeld and A. C. Kak, *Digital Picture Processing*. Academic Press, 1976.
[2] R. Szeliski, *Computer Vision: Algorithms and Applications*. Springer, 2010.
[3] D. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*. Prentice Hall, 2002.
[4] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.
[5] R. Nock and F. Nielsen, "Statistical region merging," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 11, pp. 1452–1458, 2004.
[6] N. Dhanachandra, K. Manglem, and Y. J. Chanu, "Image segmentation using K-means clustering algorithm and subtractive clustering algorithm," *Procedia Computer Science*, vol. 54, pp. 764–771, 2015.
[7] L. Najman and M. Schmitt, "Watershed of a continuous function," *Signal Processing*, vol. 38, no. 1, pp. 99–112, 1994.
[8] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331, 1988.
[9] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1222–1239, 2001.
[10] N. Plath, M. Toussaint, and S. Nakajima, "Multi-class image segmentation using conditional random fields and global classification," in *International Conference on Machine Learning*. ACM, 2009, pp. 817–824.
[11] J.-L. Starck, M. Elad, and D. L. Donoho, "Image decomposition via the combination of sparse representations and a variational approach," *IEEE Transactions on Image Processing*, vol. 14, no. 10, pp. 1570–1582, 2005.
[12] S. Minaee and Y. Wang, "An ADMM approach to masked signal decomposition using subspace representation," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3192–3204, 2019.

[13] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[14] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *arXiv preprint arXiv:2001.05566*, 2020.

[15] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner *et al.*, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[16] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, vol. 36, no. 4, pp. 193–202, 1980.

[17] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, pp. 328–339, 1989.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[21] https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

[22] D. E. Rumelhart, G. E. Hinton, R. J. Williams *et al.*, "Learning representations by back-propagating errors," *Cognitive Modeling*, vol. 5, no. 3, p. 1, 1988.

[23] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[24] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[25] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[26] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[27] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[28] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[29] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," *arXiv preprint arXiv:1701.07875*, 2017.

[30] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[31] G. Wang, W. Li, S. Ourselin, and T. Vercauteren, "Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks," in *International MICCAI Brainlesion Workshop*. Springer, 2017, pp. 178–190.

[32] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei, "Fully convolutional instance-aware semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2359–2367.

[33] Y. Yuan, M. Chao, and Y.-C. Lo, "Automatic skin lesion segmentation using deep fully convolutional networks with Jaccard distance," *IEEE Transactions on Medical Imaging*, vol. 36, no. 9, pp. 1876–1886, 2017.

[34] N. Liu, H. Li, M. Zhang, J. Liu, Z. Sun, and T. Tan, "Accurate iris segmentation in non-cooperative environments using fully convolutional networks," in *International Conference on Biometrics*. IEEE, 2016, pp. 1–8.

[35] W. Liu, A. Rabinovich, and A. C. Berg, "ParseNet: Looking wider to see better," *arXiv preprint arXiv:1506.04579*, 2015.

[36] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," *arXiv preprint arXiv:1412.7062*, 2014.

[37] A. G. Schwing and R. Urtasun, "Fully connected deep structured networks," *arXiv preprint arXiv:1503.02351*, 2015.

[38] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, "Conditional random fields as recurrent neural networks," in *IEEE International Conference on Computer Vision*, 2015, pp. 1529–1537.

[39] G. Lin, C. Shen, A. Van Den Hengel, and I. Reid, "Efficient piecewise training of deep structured models for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3194–3203.

[40] Z. Liu, X. Li, P. Luo, C.-C. Loy, and X. Tang, "Semantic image segmentation via deep parsing network," in *IEEE International Conference on Computer Vision*, 2015, pp. 1377–1385.

[41] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *IEEE International Conference on Computer Vision*, 2015, pp. 1520–1528.

[42] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," *arXiv preprint arXiv:1909.11065*, 2019.

[43] J. Fu, J. Liu, Y. Wang, J. Zhou, C. Wang, and H. Lu, "Stacked deconvolutional network for semantic segmentation," *IEEE Transactions on Image Processing*, 2019.

[44] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting encoder representations for efficient semantic segmentation," in *IEEE International Conference on Visual Communications and Image Processing*. IEEE, 2017, pp. 1–4.

[45] X. Xia and B. Kulis, "W-Net: A deep model for fully unsupervised image segmentation," *arXiv preprint arXiv:1711.08506*, 2017.

[46] Y. Cheng, R. Cai, Z. Li, X. Zhao, and K. Huang, "Locality-sensitive deconvolution networks with gated fusion for RGB-D indoor semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3029–3037.

[47] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[48] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 3–11.

[49] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018.

[50] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 424–432.

[51] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *International Conference on 3D Vision*. IEEE, 2016, pp. 565–571.

[52] T. Brosch, L. Y. Tang, Y. Yoo, D. K. Li, A. Traboulsee, and R. Tam, "Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1229–1239, 2016.

[53] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125.

[54] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2881–2890.

[55] G. Ghiasi and C. C. Fowlkes, "Laplacian pyramid reconstruction and refinement for semantic segmentation," in *European Conference on Computer Vision*. Springer, 2016, pp. 519–534.

[56] J. He, Z. Deng, and Y. Qiao, "Dynamic multi-scale filters for semantic segmentation," in *IEEE International Conference on Computer Vision*, 2019, pp. 3562–3572.

[57] H. Ding, X. Jiang, B. Shuai, A. Qun Liu, and G. Wang, "Context contrasted feature and gated multi-scale aggregation for scene segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2393–2402.

[58] J. He, Z. Deng, L. Zhou, Y. Wang, and Y. Qiao, "Adaptive pyramid context network for semantic segmentation," in *Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7519–7528.

[59] D. Lin, Y. Ji, D. Lischinski, D. Cohen-Or, and H. Huang, "Multi-scale context intertwining for semantic segmentation," in *European Conference on Computer Vision*, 2018, pp. 603–619.

[60] G. Li, Y. Xie, L. Lin, and Y. Yu, "Instance-level salient object segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2386–2395.

[61] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[62] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *IEEE International Conference on Computer Vision*, 2017, pp. 2961–2969.

[63] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8759–8768.

[64] J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3150–3158.

[65] R. Hu, P. Dollár, K. He, T. Darrell, and R. Girshick, "Learning to segment every thing," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4233–4241.

[66] L.-C. Chen, A. Hermans, G. Papandreou, F. Schroff, P. Wang, and H. Adam, "Masklab: Instance segmentation by refining object detection with semantic and direction features," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4013–4022.

[67] X. Chen, R. Girshick, K. He, and P. Dollár, "Tensormask: A foundation for dense object segmentation," *arXiv preprint arXiv:1903.12174*, 2019.

[68] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 379–387.

[69] P. O. Pinheiro, R. Collobert, and P. Dollár, "Learning to segment object candidates," in *Advances in Neural Information Processing Systems*, 2015, pp. 1990–1998.

[70] E. Xie, P. Sun, X. Song, W. Wang, X. Liu, D. Liang, C. Shen, and P. Luo, "PolarMask: Single shot instance segmentation with polar representation," *arXiv preprint arXiv:1909.13226*, 2019.

[71] Z. Hayder, X. He, and M. Salzmann, "Boundary-aware instance segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5696–5704.

[72] Y. Lee and J. Park, "CenterMask: Real-time anchor-free instance segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 906–13 915.

[73] M. Bai and R. Urtasun, "Deep watershed transform for instance segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5221–5229.

[74] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT: Real-time instance segmentation," in *IEEE International Conference on Computer Vision*, 2019, pp. 9157–9166.

[75] A. Fathi, Z. Wojna, V. Rathod, P. Wang, H. O. Song, S. Guadarrama, and K. P. Murphy, "Semantic instance segmentation via deep metric learning," *arXiv preprint arXiv:1703.10277*, 2017.

[76] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[77] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[78] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, "Understanding convolution for semantic segmentation," in *IEEE Winter Conference on Applications of Computer Vision*, 2018, pp. 1451–1460.

[79] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "DenseASPP for semantic segmentation in street scenes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3684–3692.

[80] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "ENet: A deep neural network architecture for real-time semantic segmentation," *arXiv preprint arXiv:1606.02147*, 2016.

[81] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *European Conference on Computer Vision*, 2018, pp. 801–818.

[82] F. Visin, M. Ciccone, A. Romero, K. Kastner, K. Cho, Y. Bengio, M. Matteucci, and A. Courville, "ReSeg: A recurrent neural network-based model for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 41–48.

[83] F. Visin, K. Kastner, K. Cho, M. Matteucci, A. Courville, and Y. Bengio, "ReNet: A recurrent neural network based alternative to convolutional networks," *arXiv preprint arXiv:1505.00393*, 2015.

[84] W. Byeon, T. M. Breuel, F. Raue, and M. Liwicki, "Scene labeling with LSTM recurrent neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3547–3555.

[85] X. Liang, X. Shen, J. Feng, L. Lin, and S. Yan, "Semantic object parsing with graph LSTM," in *European Conference on Computer Vision*. Springer, 2016, pp. 125–143.

[86] Y. Xiang and D. Fox, "DA-RNN: Semantic mapping with data associated recurrent neural networks," *arXiv:1703.03098*, 2017.

[87] R. Hu, M. Rohrbach, and T. Darrell, "Segmentation from natural language expressions," in *European Conference on Computer Vision*. Springer, 2016, pp. 108–124.

[88] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3640–3649.

[89] Q. Huang, C. Xia, C. Wu, S. Li, Y. Wang, Y. Song, and C.-C. J. Kuo, "Semantic segmentation with reverse attention," *arXiv preprint arXiv:1707.06426*, 2017.

[90] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," *arXiv preprint arXiv:1805.10180*, 2018.

[91] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.

[92] Y. Yuan and J. Wang, "OCNet: Object context network for scene parsing," *arXiv preprint arXiv:1809.00916*, 2018.

[93] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, Z. Zhang, H. Lin, Y. Sun, T. He, J. Mueller, R. Manmatha *et al.*, "Resnest: Split-attention networks," *arXiv preprint arXiv:2004.08955*, 2020.

[94] S. Choi, J. T. Kim, and J. Choo, "Cars can't fly up in the sky: Improving urban-scene segmentation via height-driven attention networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9373–9383.

[95] X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin, and H. Liu, "Expectation-maximization attention networks for semantic segmentation," in *IEEE International Conference on Computer Vision*, 2019, pp. 9167–9176.

[96] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *IEEE International Conference on Computer Vision*, 2019, pp. 603–612.

[97] M. Ren and R. S. Zemel, "End-to-end instance segmentation with recurrent attention," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6656–6664.

[98] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. Change Loy, D. Lin, and J. Jia, "PSANet: Point-wise spatial attention network for scene parsing," in *European Conference on Computer Vision*, 2018, pp. 267–283.

[99] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a discriminative feature network for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1857–1866.

[100] P. Luc, C. Couprie, S. Chintala, and J. Verbeek, "Semantic segmentation using adversarial networks," *arXiv preprint arXiv:1611.08408*, 2016.

[101] N. Souly, C. Spampinato, and M. Shah, "Semi supervised semantic segmentation using generative adversarial network," in *IEEE International Conference on Computer Vision*, 2017, pp. 5688–5696.

[102] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang, "Adversarial learning for semi-supervised semantic segmentation," *arXiv preprint arXiv:1802.07934*, 2018.

[103] Y. Xue, T. Xu, H. Zhang, L. R. Long, and X. Huang, "SegAN: Adversarial network with multi-scale L1 loss for medical image segmentation," *Neuroinformatics*, vol. 16, no. 3-4, pp. 383–392, 2018.

[104] M. Majurski, P. Manescu, S. Padi, N. Schaub, N. Hotaling, C. Simon Jr, and P. Bajcsy, "Cell image segmentation using generative adversarial networks, transfer learning, and augmentations," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[105] K. Ehsani, R. Mottaghi, and A. Farhadi, "SegAN: Segmenting and generating the invisible," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6144–6153.

[106] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Transactions on Image Processing*, vol. 10, no. 2, pp. 266–277, 2001.

[107] X. Chen, B. M. Williams, S. R. Vallabhaneni, G. Czanner, R. Williams, and Y. Zheng, "Learning active contour models for

medical image segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 632–11 640.

[108] S. Gur, L. Wolf, L. Golgher, and P. Blinder, "Unsupervised microvascular image segmentation using an active contours mimicking neural network," in *IEEE International Conference on Computer Vision*, 2019, pp. 10 722–10 731.

[109] P. Marquez-Neila, L. Baumela, and L. Alvarez, "A morphological approach to curvature-based evolution of curves and surfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 2–17, 2014.

[110] T. H. N. Le, K. G. Quach, K. Luu, C. N. Duong, and M. Savvides, "Reformulating level sets as deep recurrent neural network approach to semantic segmentation," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2393–2407, 2018.

[111] C. Rupprecht, E. Huaroc, M. Baust, and N. Navab, "Deep active contours," *arXiv preprint arXiv:1607.05074*, 2016.

[112] A. Hatamizadeh, A. Hoogi, D. Sengupta, W. Lu, B. Wilcox, D. Rubin, and D. Terzopoulos, "Deep active lesion segmentation," in *International Workshop on Machine Learning in Medical Imaging*, ser. Lecture Notes in Computer Science, vol. 11861. Springer, 2019, pp. 98–105.

[113] D. Marcos, D. Tuia, B. Kellenberger, L. Zhang, M. Bai, R. Liao, and R. Urtasun, "Learning deep structured active contours end-to-end," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8877–8885.

[114] D. Cheng, R. Liao, S. Fidler, and R. Urtasun, "DARNet: Deep active ray network for building segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7431–7439.

[115] A. Hatamizadeh, D. Sengupta, and D. Terzopoulos, "End-to-end trainable deep active contour models for automated image segmentation: Delineating buildings in aerial imagery," in *European Conference on Computer Vision*, 2020, pp. 730–746.

[116] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7151–7160.

[117] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1925–1934.

[118] G. Song, H. Myeong, and K. Mu Lee, "SeedNet: Automatic seed generation with deep reinforcement learning for robust interactive segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1760–1768.

[119] J. Dai, K. He, and J. Sun, "BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *IEEE International Conference on Computer Vision*, 2015, pp. 1635–1643.

[120] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters — improve semantic segmentation by global convolutional network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4353–4361.

[121] Z. Wu, C. Shen, and A. Van Den Hengel, "Wider or deeper: Revisiting the resnet model for visual recognition," *Pattern Recognition*, vol. 90, pp. 119–133, 2019.

[122] Z. Zhang, X. Zhang, C. Peng, X. Xue, and J. Sun, "ExFuse: Enhancing feature fusion for semantic segmentation," in *European Conference on Computer Vision*, 2018, pp. 269–284.

[123] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich, "Feedforward semantic segmentation with zoom-out features," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3376–3385.

[124] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3395–3402.

[125] P. Luo, G. Wang, L. Lin, and X. Wang, "Deep dual learning for semantic image segmentation," in *IEEE International Conference on Computer Vision*, 2017, pp. 2718–2726.

[126] X. Li, Z. Jie, W. Wang, C. Liu, J. Yang, X. Shen, Z. Lin, Q. Chen, S. Yan, and J. Feng, "FoveaNet: Perspective-aware urban scene parsing," in *IEEE International Conference on Computer Vision*, 2017, pp. 784–792.

[127] I. Kreso, S. Segvic, and J. Krapac, "Ladder-style densenets for semantic segmentation of large natural images," in *IEEE International Conference on Computer Vision*, 2017, pp. 238–245.

[128] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *European Conference on Computer Vision*, 2018, pp. 325–341.

[129] B. Cheng, L.-C. Chen, Y. Wei, Y. Zhu, Z. Huang, J. Xiong, T. S. Huang, W.-M. Hwu, and H. Shi, "SPGNet: Semantic prediction guidance for scene parsing," in *IEEE International Conference on Computer Vision*, 2019, pp. 5218–5228.

[130] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler, "Gated-SCNN: Gated shape cnns for semantic segmentation," in *IEEE International Conference on Computer Vision*, 2019, pp. 5229–5238.

[131] J. Fu, J. Liu, Y. Wang, Y. Li, Y. Bao, J. Tang, and H. Lu, "Adaptive context network for scene parsing," in *IEEE International Conference on Computer Vision*, 2019, pp. 6748–6757.

[132] X. Liang, H. Zhou, and E. Xing, "Dynamic-structured semantic propagation network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 752–761.

[133] X. Liang, Z. Hu, H. Zhang, L. Lin, and E. P. Xing, "Symbolic graph reasoning meets convolutions," in *Advances in Neural Information Processing Systems*, 2018, pp. 1853–1863.

[134] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ADE20K dataset," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[135] R. Zhang, S. Tang, Y. Zhang, J. Li, and S. Yan, "Scale-adaptive convolutions for scene parsing," in *IEEE International Conference on Computer Vision*, 2017, pp. 2031–2039.

[136] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *European Conference on Computer Vision*, 2018, pp. 418–434.

[137] B. Zoph, G. Ghiasi, T.-Y. Lin, Y. Cui, H. Liu, E. D. Cubuk, and Q. V. Le, "Rethinking pre-training and self-training," *arXiv preprint arXiv:2006.06882*, 2020.

[138] X. Zhang, H. Xu, H. Mo, J. Tan, C. Yang, and W. Ren, "DCNAS: Densely connected neural architecture search for semantic image segmentation," *arXiv preprint arXiv:2003.11883*, 2020.

[139] A. Tao, K. Sapra, and B. Catanzaro, "Hierarchical multi-scale attention for semantic segmentation," *arXiv preprint arXiv:2005.10821*, 2020.

[140] D. Seichter, M. Köhler, B. Lewandowski, T. Wengefeld, and H.-M. Gross, "Efficient rgb-d semantic segmentation for indoor scene analysis," *arXiv preprint arXiv:2011.06961*, 2020.

[141] M. Zhen, J. Wang, L. Zhou, S. Li, T. Shen, J. Shang, T. Fang, and L. Quan, "Joint semantic segmentation and boundary detection using iterative pyramid contexts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 666–13 675.

[142] Y. Li, L. Song, Y. Chen, Z. Li, X. Zhang, X. Wang, and J. Sun, "Learning dynamic routing for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8553–8562.

[143] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9404–9413.

[144] A. Kirillov, R. Girshick, K. He, and P. Dollar, "Panoptic feature pyramid networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6399–6408.

[145] Y. Li, X. Chen, Z. Zhu, L. Xie, G. Huang, D. Du, and X. Wang, "Attention-guided unified network for panoptic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[146] L. Porzi, S. R. Bulo, A. Colovic, and P. Kontschieder, "Seamless scene segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8277–8286.

[147] B. Cheng, M. D. Collins, Y. Zhu, T. Liu, T. S. Huang, H. Adam, and L.-C. Chen, "Panoptic-DeepLab," *arXiv preprint arXiv:1910.04751*, 2019.

[148] Y. Xiong, R. Liao, H. Zhao, R. Hu, M. Bai, E. Yumer, and R. Urtasun, "UPSNet: A unified panoptic segmentation network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8818–8826.

[149] R. Mohan and A. Valada, "EfficientPS: Efficient panoptic segmentation," *arXiv preprint arXiv:2004.02307*, 2020.

[150] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, pp. 303–338, 2010.

[151] http://host.robots.ox.ac.uk/pascal/VOC/voc2012/.

[152] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille, "The role of context for object detection

and semantic segmentation in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 891–898.

[153] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *European Conference on Computer Vision*. Springer, 2014.

[154] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3213–3223.

[155] C. Liu, J. Yuen, and A. Torralba, "Nonparametric scene parsing: Label transfer via dense scene alignment," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[156] S. Gould, R. Fulton, and D. Koller, "Decomposing a scene into geometric and semantically consistent regions," in *International Conference on Computer Vision*. IEEE, 2009, pp. 1–8.

[157] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *International Conference on Computer Vision*, vol. 2, July 2001, pp. 416–423.

[158] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari, "Learning object class detectors from weakly annotated video," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3282–3289.

[159] S. D. Jain and K. Grauman, "Supervoxel-consistent foreground propagation in video," in *European Conference on Computer Vision*. Springer, 2014, pp. 656–671.

[160] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[161] J. M. Alvarez, T. Gevers, Y. LeCun, and A. M. Lopez, "Road scene segmentation from a single image," in *European Conference on Computer Vision*. Springer, 2012, pp. 376–389.

[162] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *International Conference on Computer Vision*. IEEE, 2011, pp. 991–998.

[163] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille, "Detect what you can: Detecting and representing objects using holistic models and body parts," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1971–1978.

[164] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3234–3243.

[165] X. Shen, A. Hertzmann, J. Jia, S. Paris, B. Price, E. Shechtman, and I. Sachs, "Automatic portrait segmentation for image stylization," in *Computer Graphics Forum*, vol. 35, no. 2. Wiley Online Library, 2016, pp. 93–102.

[166] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *European Conference on Computer Vision*. Springer, 2012, pp. 746–760.

[167] J. Xiao, A. Owens, and A. Torralba, "Sun3D: A database of big spaces reconstructed using SFM and object labels," in *IEEE International Conference on Computer Vision*, 2013, pp. 1625–1632.

[168] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun RGB-D: A RGB-D scene understanding benchmark suite," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 567–576.

[169] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3D reconstructions of indoor scenes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5828–5839.

[170] I. Armeni, A. Sax, A. Zamir, and S. Savarese, "Joint 2D-3D-Semantic Data for Indoor Scene Understanding," *ArXiv e-prints*, Feb. 2017.

[171] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view RGB-D object dataset," in *IEEE International Conference on Robotics and Automation*. IEEE, 2011, pp. 1817–1824.

[172] C.-Y. Fu, M. Shvets, and A. C. Berg, "RetinaMask: Learning to predict masks improves state-of-the-art single-shot detection for free," *arXiv preprint arXiv:1901.03353*, 2019.

[173] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollár, "Learning to refine object segments," in *European Conference on Computer Vision*. Springer, 2016, pp. 75–91.

[174] H. Liu, C. Peng, C. Yu, J. Wang, X. Liu, G. Yu, and W. Jiang, "An end-to-end network for panoptic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6172–6181.

[175] K. Sofiiuk, O. Barinova, and A. Konushin, "AdaptIS: Adaptive instance selection network," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7355–7363.

[176] J. Lazarow, K. Lee, K. Shi, and Z. Tu, "Learning instance occlusion for panoptic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 720–10 729.

[177] Z. Deng, S. Todorovic, and L. Jan Latecki, "Semantic segmentation of RGBD images with mutex constraints," in *IEEE International Conference on Computer Vision*, 2015, pp. 1733–1741.

[178] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *IEEE International Conference on Computer Vision*, 2015, pp. 2650–2658.

[179] A. Mousavian, H. Pirsiavash, and J. Kosecka, "Joint semantic segmentation and depth estimation with deep convolutional networks," in *International Conference on 3D Vision*. IEEE, 2016.

[180] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," *arXiv preprint arXiv:1511.02680*, 2015.

[181] X. Qi, R. Liao, J. Jia, S. Fidler, and R. Urtasun, "3D graph neural networks for RGBD semantic segmentation," in *IEEE International Conference on Computer Vision*, 2017, pp. 5199–5208.

[182] W. Wang and U. Neumann, "Depth-aware CNN for RGB-D segmentation," in *European Conference on Computer Vision*, 2018, pp. 135–150.

[183] S. Vandenhende, S. Georgoulis, and L. Van Gool, "Mti-net: Multi-scale task interaction networks for multi-task learning," *arXiv preprint arXiv:2001.06902*, 2020.

[184] S.-J. Park, K.-S. Hong, and S. Lee, "RDFNet: RGB-D multi-level residual feature fusion for indoor semantic segmentation," in *IEEE International Conference on Computer Vision*, 2017, pp. 4980–4989.

[185] J. Jiao, Y. Wei, Z. Jie, H. Shi, R. W. Lau, and T. S. Huang, "Geometry-aware distillation for indoor semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2869–2878.

[186] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, vol. 5, no. 1, pp. 44–53, 2018.

[187] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[188] V. Goel, J. Weng, and P. Poupart, "Unsupervised video object segmentation for deep reinforcement learning," in *Advances in Neural Information Processing Systems*, 2018, pp. 5683–5694.

[189] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 152, pp. 166 – 177, 2019.

[190] L. Gao, Y. Zhang, F. Zou, J. Shao, and J. Lai, "Unsupervised urban scene segmentation via domain adaptation," *Neurocomputing*, vol. 406, pp. 295 – 301, 2020.

[191] M. Paoletti, J. Haut, J. Plaza, and A. Plaza, "Deep learning classifiers for hyperspectral imaging: A review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 158, pp. 279 – 317, 2019.

[192] J. F. Abrams, A. Vashishtha, S. T. Wong, A. Nguyen, A. Mohamed, S. Wieser, A. Kuijper, A. Wilting, and A. Mukhopadhyay, "Habitat-Net: Segmentation of habitat images using deep learning," *Ecological Informatics*, vol. 51, pp. 121 – 128, 2019.

[193] M. Kerkech, A. Hafiane, and R. Canals, "Vine disease detection in UAV multispectral images using optimized image registration and deep learning segmentation approach," *Computers and Electronics in Agriculture*, vol. 174, p. 105446, 2020.

[194] Y. Song, Z. Huang, C. Shen, H. Shi, and D. A. Lange, "Deep learning-based automated image segmentation for concrete petrographic analysis," *Cement and Concrete Research*, vol. 135, p. 106118, 2020.

[195] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Medical Image Analysis*, vol. 63, p. 101693, 2020.

[196] A. Amyar, R. Modzelewski, H. Li, and S. Ruan, "Multi-task deep learning based CT imaging analysis for COVID-19 pneumonia: Classification and segmentation," *Computers in Biology and Medicine*, vol. 126, p. 104037, 2020.

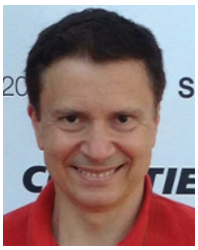
Shervin Minaee is a machine learning lead in the computer vision team at Snapchat, Inc. He received his PhD in Electrical Engineering and Computer Science from New York University, in 2018. His research interests include computer vision, image segmentation, biometric recognition, and applied deep learning. He has published more than 40 papers and patents during his PhD. He previously worked as a research scientist at Samsung Research, AT&T Labs, Huawei Labs, and as a data scientist at Expedia group. He has been a reviewer for more than 20 computer vision related journals from IEEE, ACM, Elsevier, and Springer. He has won several awards, including the best research presentation at Samsung Research America in 2017 and the Verizon Open Innovation Challenge Award in 2016.

Yuri Boykov is a Professor in the Cheriton School of Computer Science at the University of Waterloo. His research is concentrated in the area of computer vision and biomedical image analysis with focus on modeling and optimization for structured segmentation, restoration, registration, stereo, motion, model fitting, recognition, photo-video editing and other data analysis problems. He is an editor for the International Journal of Computer Vision (IJCV). His work was listed among the 10 most influential papers in IEEE TPAMI (Top Picks for 30 years). In 2017 Google Scholar listed his work on segmentation as a "classic paper in computer vision and pattern recognition" (from 2006). In 2011 he received the Helmholtz Prize from the IEEE and the Test of Time Award from the *International Conference on Computer Vision*.

Fatih Porikli is an IEEE Fellow and a Senior Director at Qualcomm, San Diego. He was a full Professor in the Research School of Engineering at the Australian National University and, until recently, a Vice President at Huawei CBG Device; Hardware, San Diego. He led the Computer Vision Research Group at NICTA, Australia, and was a Distinguished Research Scientist at Mitsubishi Electric Research Laboratories, Cambridge, MA. He received his PhD degree from New York University in 2002. He was the recipient of the R&D 100 Scientist of the Year Award in 2006. He won six best paper awards, authored more than 250 papers, co-edited two books, and invented over 100 patents. He has served as the General Chair and Technical Program Chair of many IEEE conferences and as an Associate Editor of premier IEEE and Springer journals for the past 15 years.

Antonio Plaza is an IEEE Fellow and a professor in the Department of Technology of Computers and Communications at the University of Extremadura, where he received the MSc degree in 1999 and the PhD degree in 2002, both in Computer Engineering. He has authored more than 600 publications, including 300 JCR journal papers (more than 170 in IEEE journals), 24 book chapters, and over 300 peer-reviewed conference proceedings papers. He is a recipient of the Best Column Award of the IEEE Signal Processing Magazine in 2015, the 2013 Best Paper Award of the *JSTARS* journal, and the most highly cited paper (2005–2010) in the *Journal of Parallel and Distributed Computing*. He is included in the 2018, 2019 and 2020 Highly Cited Researchers List.

Nasser Kehtarnavaz is a Distinguished Professor in the Department of Electrical and Computer Engineering at the University of Texas at Dallas, Richardson, TX. His research interests include signal and image processing, machine learning, and real-time implementation on embedded processors. He has authored or co-authored ten books and more than 390 journal papers, conference papers, patents, manuals, and editorials in these areas. He is a Fellow of the SPIE, a licensed Professional Engineer, and Editor-in-Chief of the *Journal of Real-Time Image Processing*.



Demetri Terzopoulos is a Distinguished Professor and Chancellor's Professor of Computer Science at the University of California, Los Angeles, where he directs the UCLA Computer Graphics & Vision Laboratory, and is Co-Founder and Chief Scientist of VoxelCloud, Inc. He received his PhD degree in Artificial Intelligence from the Massachusetts Institute of Technology (MIT) in 1984. He is or was a Guggenheim Fellow, a Fellow of the ACM, IEEE, IETI, Royal Society of Canada, and Royal Society of London, and a Member of the European Academy of Sciences, the New York Academy of Sciences, and Sigma Xi. Among his many awards are an Academy Award from the Academy of Motion Picture Arts and Sciences for his pioneering work on physics-based computer animation, and the Computer Pioneer Award, Helmholtz Prize, and inaugural Computer Vision Distinguished Researcher Award from the IEEE for his pioneering and sustained research on deformable models and their applications. *Deformable models*, a term he coined, is listed in the IEEE Taxonomy.