Neurocomputing 338 (2019) 321–348

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

# Survey on semantic segmentation using deep learning techniques

Fahad Lateef*, Yassine Ruichek

*LE2I-CNRS, University of Technology of Belfort-Montbeliard (UTBM), France*

A R T I C L E I N F O

*Article history:*
Received 26 October 2018
Revised 31 December 2018
Accepted 1 February 2019
Available online 8 February 2019

Communicated by Haijun Jiang

*Keywords:*
Deep learning
Semantic segmentation
Recurrent neural network
Semi-weakly supervised networks

A B S T R A C T

Semantic segmentation is a challenging task in computer vision systems. A lot of methods have been developed to tackle this problem ranging from autonomous vehicles, human-computer interaction, to robotics, medical research, agriculture and so on. Many of these methods have been built using the deep learning paradigm that has shown a salient performance. For this reason, we propose to survey these methods by, first categorizing them into ten different classes according to the common concepts underlying their architectures. Second, by providing an overview of the publicly available datasets on which they have been assessed. In addition, we present the common evaluation matrix used to measure their accuracy. Moreover, we focus on some of the methods and look closely at their architectures in order to find out how they have achieved their reported performances. Finally, we conclude by discussing some of the open problems and their possible solutions.



## 1. Introduction

The recent works in deep learning dealing with semantic segmentation have been significantly improved by using neural networks. Neural networks have a long history since the 1940s and they did not get much of the attention of researchers until 1990s [1]. Neutral networks made huge progress because of large amount of data is available thanks to the rise of digital cameras, cell phone cameras, and the computing power, which is getting faster as GPUs become general purpose computing tools.

Deep neural networks are very effective in semantic segmentation, that is labeling each region or pixel with a class of objects/non-objects. Semantic segmentation plays an important role in image understanding and essential for image analysis tasks. It has several applications in computer vision & artificial intelligence - autonomous driving [2,3], robot navigation [4], industrial inspection [5]; remote sensing [6]; In cognitive and computational sciences - saliency object detection [7,8]; In Agriculture sciences [9]; Fashion - categorizing clothing items [10]; In medical sciences - medical imaging analysis [11] etc. The earlier approaches used for semantic segmentation were textonforest [12], random-forest based classifiers [13], whereas deep learning techniques allowed precise and much faster segmentation [14].

Semantic segmentation requires image classification, object detection, and boundary localization. Fig. 1 is an example of object detection, involving bounding box, and classification of each pixel into different classes (car, road, sky, vegetation, terrain etc).

Deep learning is a new field division of machine learning, which is rapidly growing with the pace making it very difficult to stay up to date, even to keep track of the works dealing with semantic segmentation. These works cover the development of new methods, improvements of existing methods, and their deployment in new application domains. This is the reason that there is a lack of state-of-the-art reviews.

Some surveys and review papers have addressed advancements and innovations on the subject of deep learning and semantic segmentation. A Survey by Zhu et al. [15] covering a wide range of the papers and areas of semantic segmentation topics including, interactive methods, recent development in the super-pixel, object proposals, semantic image parsing, image co-segmentation, semi & weakly supervised, and fully supervised image segmentation. Thoma [16] presented a taxonomy of segmentation algorithms and overview of completely automatic, passive, semantic segmentation algorithms. Niemeijer et al. [17] presented a review of neural network based semantic segmentation for scene understanding in the context of the autonomous driving. Guo et al. [18] provided a review of semantic segmentation approaches, i.e., region-based, FCN-based and weakly supervised approaches. They have summarized the strengths, weaknesses and major challenges in image semantic segmentation. Geng et al. [19] presented a survey of recent progress in semantic segmentation with CNN's, and newly developed strategies that have achieved promising results on the Pascal VOC 2012 semantic segmentation challenge. Detail review provided by Garcia-Garcia et al. [20] on deep learning methods for semantic

* Corresponding author.
*E-mail addresses:* fahad.lateef@utbm.fr (F. Lateef), yassine.ruichek@utbm.fr (Y. Ruichek).

https://doi.org/10.1016/j.neucom.2019.02.003
0925-2312/

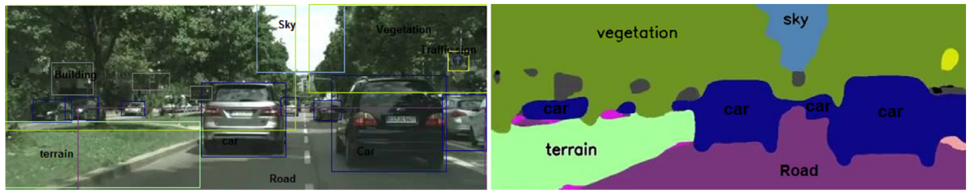



**Fig. 1.** Object detection / bounding box and semantic segmentation.

segmentation with their contributions and significance in the field. An extensive review presented by Hongshan et al. [21], categorized different methods based on hand engineered features, learned features, and weakly supervised learning.

The main contributions of this paper are as follows:

1. The existing methods have been categorized into ten different classes according to the common concept that underlays their architectures. This categorization gives a complete summary of the methods both inspire and diverge from one another.
2. More than 100 different models and 33 datasets (publicly available) have been covered, stating the corpus, original architecture, testing benchmark of each model, and the attributes of each dataset. Furthermore, we provide the best performing method yielded top classification accuracy on each dataset until date.
3. An emphasis on how these methods achieved their accomplishments is given by analyzing their structural design and their performance on the assessed datasets.
4. Finally, some of the open problems and possible solutions have been discussed.

In this survey, all the models are carefully chosen and put into relation to each other according to their architectural design and contribution to the field. This includes improving accuracy, reducing computation complexity, developing new methods, and enhancing existing ones. All the results reported in this paper are taken from the original papers. We have tried to cover most of the works in deep neural networks for semantic segmentation. This survey will help the new researchers to strengthen their understanding of these remarkable works.

## 2. Deep learning architectures for semantic segmentation

This section provides the details of all the reviewed semantic segmentation methods. We have categorized these methods into ten (10) classes, presented in the tabular form stating each method, its main idea, its architecture origin, testing benchmarks, publication date, and code availability (Table 13 provides links of available source codes).

The recent success of deep convolutional neural networks (CNNs) has enabled outstanding progress in semantic segmentation. The first successful application of convolutional neural network was developed by LeCun et al. [22]. They introduced an architecture named LeNet5 to read zip codes, digits, and extract features at multiple locations in the image. Later, Alex Krizhevsky released a large deep convolutional neural network (AlexNet) [23], which is regarded as one of the most influential publications in the field. AlexNet is a deeper and wider version of the LeNet, used to learn complex objects and object hierarchies. Zeiler and Fergus [24] presented the ZFNet, which is a fine-tuning of the AlexNet structure. They proposed a technique of visualizing feature maps at any layer in the network model. This technique uses a multi-layered deconvolutional network to project the feature activations back to the input pixel space. Lin et al. [25] proposed a Network-In-Network model based on micro neural networks, which is a multilayer perceptron (MLP) [26], consisting of multiple fully connected layers with nonlinear activation functions. Szegedy et al. [27] proposed an efficient deep neural network called **GoogleNet**. They introduced an inception module as shown in Fig. 2, which is a combination of $1 \times 1$, $3 \times 3$, and $5 \times 5$ convolutional filters and a pooling layer. It reduced the number of features and operations at each layer thus saving the time and computational cost.

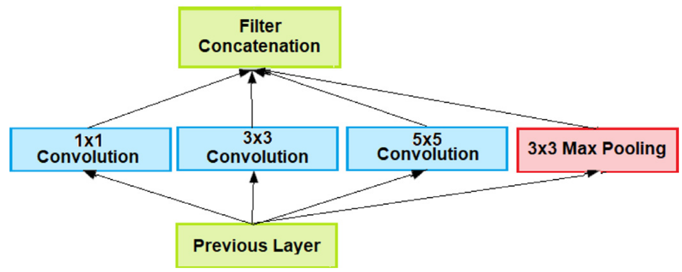



**Fig. 2.** Inception module.

The same authors proposed in [28] an algorithm refereed as **BN-Inception** for constructing, training, and performing inference with Batch Normalization method. Szegedy et al. [29] further introduced two new modules **Inception V2** and **Inception V3** with some major modifications (i.e., factorizing convolutions and using grid reduction technique) of their previous module. Later, Szegedy et al. [30] replaced the filter concatenation stage of the Inception architecture with residual connections in order to increase efficiency and performance. They proposed Inception-ResNet-v1, Inception-ResNet-v2 and a pure Inception variant called Inception V4. Chollet et al. [31] proposed a module named **Xception**, meaning extreme inception. They replaced the inception modules with depth wise separable convolutions proposed in [32]. Table 1 shows GoogLeNet Modules.

### 2.1. Feature encoder based methods

VGG [33] and ResNet [34] methods are the most dominant approaches for feature extraction. In this category, we review these methods and their invariants presented in Table 2. The idea behind the concept is to extract feature maps based on stacked convolution layers, ReLu layers and pooling layers.

#### 2.1.1. VGG network

VGG network [33] introduced by Oxford's renowned Visual Geometry Group. Unlike LeNet [22] and AlexNet [23], VGGNet uses multiple $3 \times 3$ convolution in the sequence that can match the effect of larger receptive fields, e.g. $5 \times 5$ and $7 \times 7$. However, it required a large number of parameters and learning power due to having large classifiers. Fig. 3 shows a VGGNet with 16 convolutional layers.

#### 2.1.2. Residual learning frameworks

Residual learning frameworks include methods which use residual block [34] as a fundamental building block in their architecture.

Residual Network - ResNet [34] is the most popular and widely used neural network for semantic segmentation. It is hard to train

**Table 1**
GoogLeNet modules.

| Model | Corpus | Original architecture | Testing benchmark | Published on | Code available |
|---|---|---|---|---|---|
| GoogLeNet | Inception module: Bottleneck [27] | NIN | ImageNet | September 17, 2014 | Yes |
| | Batch Normalization Modified BN-Inception [28] | Inception | ImageNet | March 2, 2015 | Yes |
| | Inception V2, V3 [29] | BN-Inception | ImageNet | December 11, 2015 | Yes |
| | Inception V4 and Inception-ResNet-v1, 2 Combining the Inception architecture with Residual connections [30] | Inception V3 ResNet | ImageNet | August 23, 2016 | Yes |
| | Xception [31] Depthwise Separable Convolutions [32] | Inception V3 ResNet | ImageNet JFT (Google's) FastEval14k | April 4, 2017 | Yes |

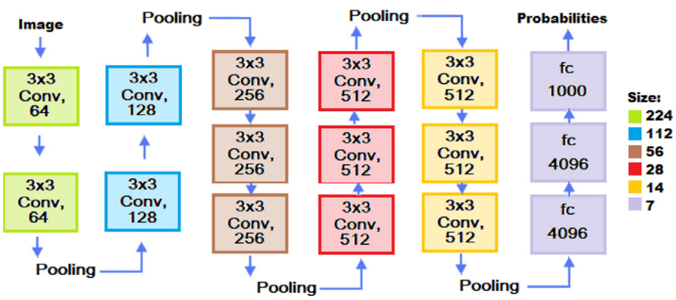



**Fig. 3.** VGG-16 layer structure.

a deep neural network with large numbers of layers, the more increase in depth, its performance gets saturated or even starts degrading due to vanishing gradient problem. Several solutions were proposed in [35–37] but none of them seemed to really tackle the problem. He et al. [34] resolved the vanishing gradient problem in an effective way by introducing identity shortcut connection (i.e., skipping one or more layers) as shown in Fig. 4. They proposed a pre-activation variant residual block in which the gradients can easily flow through the shortcut connection without obstruction during the back pass of back propagation.

Several architectures are based on ResNet, its variants and interpretations. Paszke et al. [45] presented an encoder/decoder scheme network called efficient neural network (ENet). This network is similar to the ResNet bottleneck approach, created specifically for tasks requiring low latency operation, i.e., mobile phones or battery-powered devices. In [49,50], the authors proposed counter-intuitive way of training a deep network by randomly dropping its layers and using the full network in testing time. Wu et al. [38] presented a neural network called ResNet-38, in which they added and removed layers in residual networks at train/test time. They analyzed the effective depths of residual units, and point out that ResNet behaves as linear ensembles of shallow networks. Pohlen et al. [44] proposed a full-resolution residual network (FRRN) with strong localization and recognition performance for semantic segmentation. FRRN exhibits the same superior training properties as ResNet, having two processing streams: residual and pooling. Residual stream carries information at the full image resolution and enables precise adherence to segment boundaries. The pooling stream undergoes a sequence of pooling operations to obtain robust features for recognition. The two streams are coupled at the full image resolution using residuals in order to realize strong recognition and localization performance for semantic segmentation. Xie et al. [41] proposed a modified ResNet called ResNeXt, following the split-transform-merge strategy as inception modules [27,30], except the outputs of different paths are concatenated and all paths share the same topology. Thus, this allows the design to extend to any large number of transformations. Adapting the idea of ResNet-50 [34], an architecture called Adaptive network or AdapNet is proposed by Valada et al. [40]. They introduced an

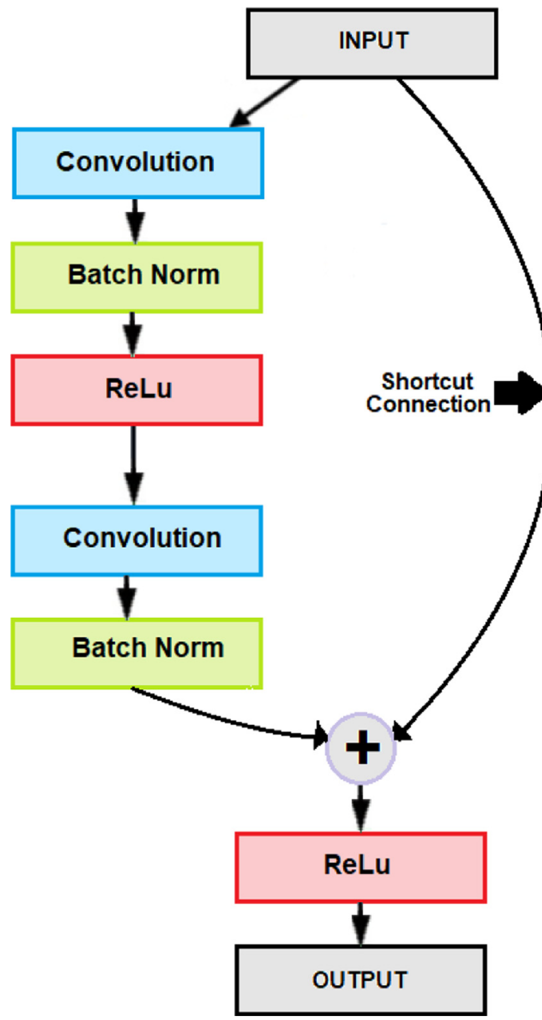



**Fig. 4.** Residual learning: a building block.

additional convolution with a kernel size of $3 \times 3$ before the first convolution layer in ResNet, which enables the network to learn more high resolution features in less time. They also proposed the convoluted mixture of deep experts (CMoDE) fusion scheme for learning robust kernels from complementary modalities and spectra. The proposed model adaptively weighs class-specific features based on the scene condition. Inspired by ENet, Romera et al. [46] proposed an efficient residual factorized network ERFNet for real-time semantic segmentation. ERFNet proposes a non-bottleneck-1D (non-bt-1D) layer and combines with bottleneck designs in a way that best leverages their learning performance and efficiency. Considering ERFNet as the baseline, Deng et al. [48] proposed a new encoder decoder approach called Restricted

**Table 2**
Feature encoder based methods.

| Category | Strategy / Structure | | Corpus | Original Architecture | Testing Benchmark | Published on | Code Available |
|---|---|---|---|---|---|---|---|
| Feature Encoder | Visual Geometry Group Network | (VGGNet) [33] | Convolutional Networks (ConvNets) Used much smaller $3 \times 3$ filters in each convolutional layers which match the effect of larger receptive fields e.g. $5 \times 5$ and $7 \times 7$. | AlexNet | ImageNet, PASCAL VOC | April 10, 2015 | YES |
| | RESIDUAL . LEARNING | Residual Network (ResNet) [34] | Bottleneck Approach Shortcut Connections are added (MLPs-Multi Layer Perceptions) | VGG | ImageNet, Cityscapes, CIFAR-10, COCO, PASCAL VOC | December 10, 2015 | YES |
| | RESIDUAL . LEARNING | ResNet-38 [38] | (Shallow Network) ReNet for Image classification FCN for semantic image segmentation. | ResNet + FCN | Cityscapes, ADE20K, PASCAL VOC | November 30, 2016 | YES |
| | RESIDUAL . LEARNING | Fully Convolutional Dense ResNet (FC-DRN) [39] | Combining the strength of FC-ResNet: gradient flow and iterative refinement. FC-DenseNet: Multi-Scale feature representation and deep supervision). | ResNet | CamVid | April 30, 2018 | - |
| | RESIDUAL . LEARNING | Adaptive Network (AdapNet) [40] | Convoluted Mixture of Deep Experts (CMoDE) fusion scheme. | ResNet | Cityscapes, Synthia, Freiburg forest | May 29, 2017 | - |
| | RESIDUAL . LEARNING | ResNeXt [41] | Hyper-parameter ”Cardinality” a new way to adjust models capacity. | ResNet | ImageNet, COCO, CIFAR | April 11, 2017 | YES |
| | RESIDUAL . LEARNING | INPLACE-ABN [42] | In-Place Activated Batch Normalization module: To reduce the training memory footprint of residual networks. | DeepLabV3 | COCO-Stuff, Cityscapes Mapillary Vistas | December 11, 2017 | YES |
| | RESIDUAL . LEARNING | Dynamic-Structured Semantic Propagation Network (DSSPN) [43] | DSSPN explicitly constructs a semantic neuron graph network by incorporating the semantic concept hierarchy. | - | ADE20K, COCO-Stuff, Cityscape Mapillary | March 16, 2018 | - |
| | FRAMEWORKS / Encoder Decoder | Full-resolution Residual Networks (FRRN) [44] | Two Stream Network Residual Stream: Carries information at the full image resolution, enabling precise adherence to segment boundaries. Pooling Stream: Sequence of pooling operations to obtain robust features for recognition. | ResNet + VGG | Cityscapes | December 6, 2016 | YES |
| | FRAMEWORKS / Encoder Decoder / Real-Time Segmentation | Efficient Neural Network (ENet) [45] | Presents a different view on encoder-decoder architecture The decoder is to upsample the output of the encoder, only to fine-tuning. | ResNet | Cityscapes, CamVid, SUN | June 7, 2016 | YES |
| | FRAMEWORKS / Encoder Decoder / Real-Time Segmentation | Efficient Residual Factorized Network (ERFNet) [46] | A non-bottleneck-1D (non-bt-1D) layer and combines with bottleneck. | ResNet ENet | Cityscapes | January 1, 2018 | YES |
| | FRAMEWORKS / Encoder Decoder / Real-Time Segmentation | Efficient Spatial Pyramid ESPNet [47] | Efficient spatial pyramid (ESP) modules: Spatial pyramid of dilated convolutions. | ResNet | CityScapes, PASCAL VOC, Mapillary | March 22, 2018 | YES |
| | FRAMEWORKS / Encoder Decoder / Real-Time Segmentation | Restricted Deformable Convolution (RDC) Network [48] | Zoom Augmentation method: Transforming conventional images to fish-eye images. | ERFNet | CityScapes, SYNTHIA | January 3, 2018 | - |

Deformable Convolution (RDC) for road scene semantic segmentation handling large distorted images. It can model geometric transformations by learning the shapes of convolutional filters conditioned on the input feature map. They proposed zoom augmentation method to convert standard images to fisheye images. Mehta et al. [47] proposed a convolutional module called efficient spatial pyramid (ESP) to their new efficient neural network. The ESP module consists of point-wise convolutions (reducing the complexity) and the spatial pyramid of dilated convolutions (providing large receptive field). Recently, Casanova et al. [39] proposed a Fully

**Table 3**
Region proposal based methods.

| Category | Strategy / Structure | Corpus | Original architecture | Testing benchmark | Published on | Code available |
|---|---|---|---|---|---|---|
| Regional Proposals | Regional Convolutional Neural Network (R-CNN) [52] | Regional proposal generator: Selective Search Method CNN: for extracting features from each region Set of class specific linear SVMs to score features. | AlexNet VGG-16 | PASCAL VOC | October 22, 2014 | Yes |
| | Fast R-CNN [54] | Improvement in R-CNN Region of Interest (RoI) pooling layer. | VGG-16 | PASCAL VOC | September 27, 2015 | Yes |
| | Faster R-CNN [55] | Region Proposal Network (RPN) Merge of RPN and Fast R-CNN. | VGG-16 FCN as RPN ZFNet | PASCAL VOC COCO | June 6, 2016 | Yes |
| | Mask R-CNN [56] | Region of Interest Alignment (RoIAlign): for pixel-to-pixel alignment | VGG-16 FCN as RPN ZFNet | Cityscapes, COCO | January 24, 2018 | Yes |
| | Feature Pyramid Network (FPN) [57] | Create feature pyramids having semantics at all levels, that can be used to replace featured image pyramids. | Fast/Faster R-CNN | COCO | April 18, 2017 | Yes |
| | Path Aggregation Network (PANet) [58] | Bottom up path augmentation adaptive feature pooling: Fully connected fusion: | Mask R-CNN / FPN | COCO, Cityscapes, Mapillary vistas | March 5, 2018 | – |

Convolutional Dense ResNet called FC-DRN. The basic idea is to combine the strength of the network architectures FC-ResNet (gradient flow and iterative refinement) and FC-DenseNet [51] (multi-scale feature representation and deep supervision). Liang et al. [43] proposed a Dynamic Structured Semantic Propagation Network (DSSPN), that builds a large semantic neuron graph by taking in the semantic concept hierarchy into network construction. In semantic propagation graph, each neuron is responsible for segmenting out regions of one concept in the word hierarchy. They proposed dense semantic-enhanced neural block in which the learned features of each neuron are further propagated into its child neurons to evolve features for recognizing more fine-grained concepts. Bulò et al. [42] present In-Place Activated Batch Normalization (INPLACE-ABN) architecture module to reduce the training memory footprint of residual network ResNeXt [41] and ResNet-38 [38].

The focus on VGG and ResNet approaches of recent works led to remarkable results in semantic segmentation. The residual learning frameworks follow the core idea “skip connection” which is the main intuition behind their success. However, using it in large scale can lead to memory problem. These groundbreaking works make it possible to train deeper networks with good performance.

### 2.2. Regional proposal based methods

Regional proposal algorithms are very influential in computer vision (for object detection techniques). The core idea is to detect the regions according to the variety of color spaces and similarity metrics, and then perform the classification (region proposals that might contain object) often called Region-wise prediction. Regional Convolutional Neural Network (R-CNN) along with its descendants shown in Table 3.

Girshick et al. [52] at UC Berkeley proposed a first region-based convolutional neural network (R-CNN) for object detection tasks. The R-CNN consists of three modules; regional proposal generator in which they used selective search method [53] performing the function of generating 2000 different regions that have the highest probability of containing an object; convolutional neural network [22] for extracting features from each region; finally these feature from CNN are used as input to set of class specific linear SVMs. The features are also fed into bounding box regressor to obtain the most accurate coordinates and reduce localization errors. Fig. 5 shows R-CNN architecture.

In [54], the authors proposed Fast R-CNN in which, a technique called RoIPool (Region of Interest Pooling) is used that improves

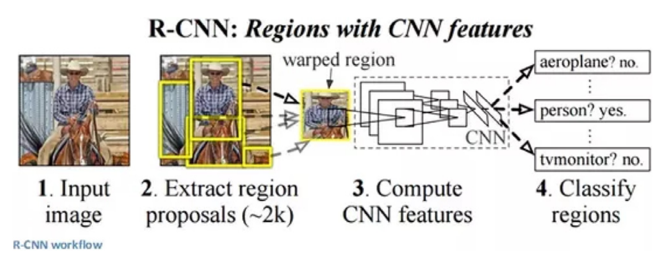



**Fig. 5.** The architecture of R-CNN [52].

the training and testing speed and increases the accuracy for object detection. Later a team from Microsoft proposed an Faster-RCNN [55] architecture. They introduce Region Proposal Network (RPN) which is a kind of fully convolutional network (FCN) constructed by adding a few additional convolutional layers that predict object bounds and objectness (set of object classes vs. background) scores at each position. The RPN generates region proposals (multiple scales and aspect ratios), which are fed into Fast R-CNN for object detection. RPN and Fast R-CNN share their convolutional features which reduce the complexity, increases the speed and overall object detection accuracy. Lin et al. [57] present Feature Pyramid Networks (FPN), a multi-scale pyramidal hierarchy of deep convolutional network (ConvNet's), and creates feature pyramids having semantics at all levels, that can be used to replace featurized image pyramids with minimal cost (power, speed, or memory). He et al. [56] proposed a Mask Regional Convolutional Neural Network (Mask-RCNN), extending Faster R-CNN to pixel-level image segmentation. It added a branch (small FCN) on each RoI for predicting object mask in a pixel-to pixel manner, in parallel with the existing branch for bounding box recognition (classification and regression). Faster R-CNN has a drawback of misalignment (pixel-to-pixel alignment) between network inputs and outputs. Mask-RCNN fixes this issue by replacing the RoI pooling layer with Region of Interest Alignment (RoIAlign), a quantization-free layer that preserves exact spatial locations as shown in Fig. 6. Recently, Liu et al. [58] proposed network built on Mask-RCNN and FPN named Path Aggregation Network (PANet), boosting information flow in proposal-based instance segmentation framework.

Region proposal based neural networks have the advantage that object detection and segmentation can be achieved at the same time. Proposals are generated by algorithms ([59] provide an in-depth analysis) that are semantically meaningful and related to

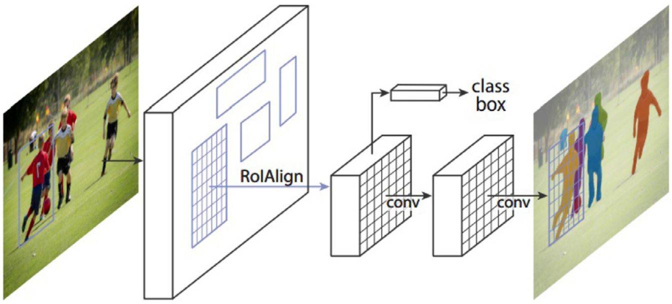



**Fig. 6.** The framework of Mask R-CNN [56].

objects. It may contain an object class or several other classes that can help in determining the semantic labels. Furthermore, feeding the wrapped region proposals into a convolutional neural network for classification can reduce the computational cost.

### 2.3. Recurrent neural network based methods

Recurrent neural networks (RNNs) were genuinely introduced for dealing sequences [60–62]. Beside its accomplishments in handwriting and speech recognition, RNNs are very much successful in computer vision tasks (dealing with images). We have only reviewed network models that adopt RNN in 2D images (integrate the convolution layers with RNNs). The Recurrent neural network made up of Long- Short-Term Memory (LSTM) [63] blocks. RNN capability to learn long term dependencies from sequential data and ability to keep memory along the sequence makes it applicable in many computer vision tasks including semantic segmentation [64,65] and scene segmentation and labeling [66,67], based on using RNN CNN combination. Table 4 shows RNN based methods.

Pinheiro and Collobert [68] proposed a convolutional neural network, which relies on a recurrent architecture ($_{R}$**CNN**). $_{R}$CNN is a sequence of shallow networks sharing same weights, at each instance using the downscaled input image and prediction maps from the previous instance of the network, and automatically learns to smooth its predicted labels. Fan et al. [66] proposed the contextual RNNs for scene labeling. The proposed network can capture long-range dependencies (GIST, local and global features) in an image. These features (after upsampling) are fused via an attention model [67]. Salvador et al. [75] present an encoder/decoder based recurrent neural network architecture for semantic instance segmentation. The proposed architecture much resembled FCN [77] architecture (encoder: feature extractor) using skip-connection, except with decoder part that is the recurrent network (convolutional LSTM [76]), predicting one instance (object in the image) at a time and output them. Byeon et al. [37] present a two-dimensional (2D) long-short term memory (LSTM) recurrent neural network for scene labeling. 2D LSTM network architecture consists of four LSTM blocks (it propagates surrounding contexts) and a feed forward layer (summing LSTM activations). This method is able to model long-range dependencies (both local and global) in image. Visin et al. [64] propose an RNN-based architecture for semantic segmentation codenamed **ReSeg** to model structured information of local generic features extracted from CNNs. The proposed model is a modified and extended version of ReNet [65]. The proposed recurrent layer is composed of multiple RNNs [73,74] that sweep the image in both directions horizontally and vertically (output of hidden states), encoding local features, and providing relevant global information. ReNet layers are stacked on top of the output of a FCN. Fig. 7 shows Reseg network architecture.

Shuai et al. [69] use graphical RNNs (Directed Acyclic Graph-Recurrent Neural Network or **DAG-RNN**) to model long-range contextual dependencies of local features in the image for semantic segmentation. They proposed a new class weighting function in order to improve the accuracy for recognition of non-frequent classes. Inspired by DenseNet [71], Fan and Ling [70] proposed a DAG structured dense Recurrent Neural Network (DD-RNNs) architecture to model vast dependencies in images through dense connections. Recently, Shuai et al. [72] proposed a DAG-RNN network to model long-range semantic dependencies for graphical structured images (DAG-structured). Their proposed segmentation network consists of three modules: Local region representation (using pre-trained CNN), context aggregation (using DAG-RNN), and feature map upsampling (deconvolution network). They also proposed a class-weighted loss during training in order to overcome class imbalance issue or give attention to rare classes.

Recurrent neural network (RNN) can be very beneficial in semantic segmentation; it has recurrent connections (ability to retain previous information) and ability to capture context in an image by modeling long-range semantic dependencies for the image.

### 2.4. Upsampling / Deconvolution based methods

Convolution neural network models have the ability to learn automatically high-level features via a layer-to-layer propagation with sacrificing the spatial information. One deep understanding is that spatial information lost during downsampling operation can be regained by upsampling and deconvolution. Second is to develop reconstruction technique for increasing spatial accuracy and refinement technique for fusing the features of a low and high level. Table 5 shows Upsampling / Deconvolution based methods.

Noh et al. [79] used this idea and developed a network model by learning a deconvolution network. The convolution network reduces the size of activations through feed forwarding, and deconvolution network enlarges the activations through the combination of unpooling and deconvolution operations. Wang et al. [78] proposed an objectness-aware semantic segmentation framework (OA-Seg) using two networks, object proposal network (OPN), predicting object bounding boxes and their objectness scores, and lightweight deconvolutional neural network (Light-DCNN) for upsampling the feature maps to larger resolution. Long et al. [77] proposed first Fully Convolutional Network (FCN), and made breakthroughs in deep learning based semantic segmentation. FCN architectures have become the standard in semantic segmentation; most of the methods utilize FCN architecture. FCN coverts the classification network [23,27,33] into fully convolutional network and produces a probability map for input of arbitrary size. FCN recovers the spatial information from the downsampling layers by adding upsampling layers to standard convolution network. They defined a skip architecture (shallow fine layer) that combines semantic information from a deep coarse layer with appearance information to produce precise and in depth segmentation. The basic idea was to re-architect and fine-tune classification model (image classification) to learn efficiently from whole image inputs and whole image ground truths (prediction of semantic segmentation). This allows hence extending these classification models to segmentation, and improving the architecture with multi-resolution layer combinations. Fig. 8 shows FCN architecture.

Badrinarayanan et al. [80] present an encoder decoder structure deep fully convolutional neural network called SegNet. The encoder network has the same topology as VGG [33] with no fully connected layers followed by a decoder network (from [93]) for a pixel-wise classification. SegNet obtains higher resolution than that in [77] by using set of decoders, each one corresponding to each encoder. One key feature of SegNet is that the information transfer is direct instead of convolving them. SegNet was one the best model to use when dealing with image segmentation problems specially scene segmentation tasks.

**Table 4**
Recurrent neural network based methods.

| Category | Strategy / Structure | | Corpus | Original architecture | Testing benchmark | Published on | Code available |
|---|---|---|---|---|---|---|---|
| Recurrent neural network | Recurrent Convolution Neural Network ($_R$CNN) [68] | | Feed-forward approach: Models non-local class dependencies in a scene from the raw image (Extract contextual information). | LeNet | Stanford Background SIFT Flow | June 21, 2014 | – |
| | Directed Acyclic Graph RNNs | DAG-RNNs [69] | Model the contextual dependencies of local features. Class weighting function that attends to rare classes. | VGGNet + RNN | SiftFlow, CamVid, Barcelona | November 23, 2015 | – |
| | | Dense Recurrent Neural Network (DD-RNN) [70] | Model contextual dependencies through dense connections inspired by DenseNet [71]. Attention model to focus on relevant dependencies. | VGGNet + RNN | PASCAL Context, ADE20K, SiftFlow | January 23, 2018 | – |
| | | DAG-RNNs [72] | Model long-range semantic dependencies for graphical structured images. Class weighting function that attends to rare classes. | VGGNet + RNN | Sift Flow, Pascal Context COCO Stuff | June 6, 2017 | – |
| | ReSeg: Recurrent Segmentation [64] | | Modified ReNet [65] Recurrent Layer: Composed by multiple RNNs. Gated Recurrent Unit (GRU) [73] or LSTM [74] | ReNet + RNN | CamVid, Oxford Flower, Weizmann Horse | June 1, 2016 | Yes |
| | Multi-level Contextual Recurrent Neural Networks (MCRNNs) [66] | | CRNNs encode three contextual cues (local, global and GIST). Attention model is adopted to improve effectiveness. | VGGNet + RNN | CamVid, KITTI, SiftFlow, Stanford-background, Cityscapes | January 23, 2018 | – |
| | Two-Dimensional LSTM Network (2D-LSTM) [37] | | Model long-range dependencies (Local: Pixel-by -Pixel and Global: Label-by-Label) in an image. LSTM blocks: Activation (surrounding contexts in all directions). Feedforward layer: Summing LSTM activations. | LSTM | Stanford Background SIFT Flow | June 7, 2015 | – |
| | Recurrent model for semantic instance segmentation [75] | | Encoder/Decoder based Recurrent Neural Network Encoder: Feature extractor Decoder: Convolutional LSTM [76], predicting one instance at a time | ResNet + Convolutional LSTM | Pascal VOC 2012, Cityscapes, CVPPP Plant Leaf Segmentation | March 22, 2018 | Yes |

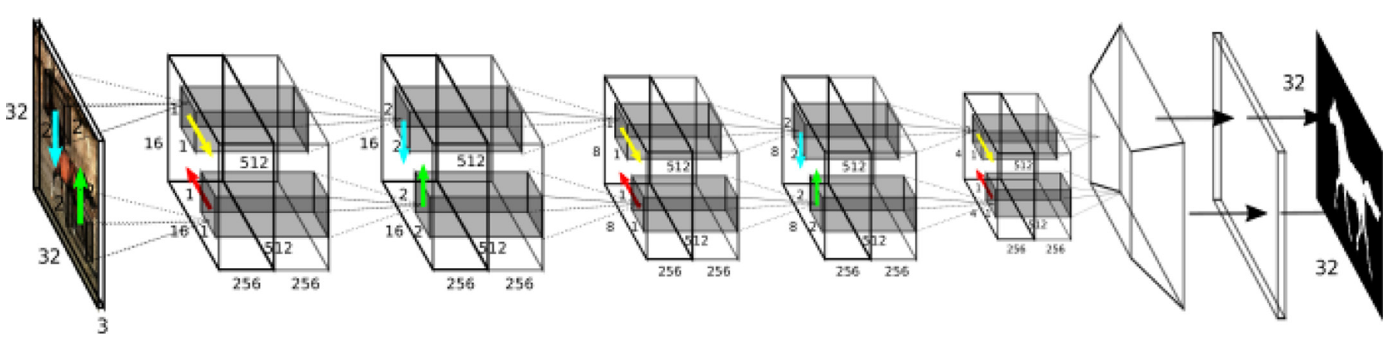


**Fig. 7.** ReSeg network [64].

Ghiasi and Fowlkes [92] proposed a network called the Laplacian Pyramid Reconstruction and Refinement (LRR) since the architecture uses a Laplacian reconstruction pyramid [94]. The architecture uses low-resolution feature maps to reconstruct a coarse and low frequency segmentation map, and then refines this map by adding in higher frequency details derived from higher-resolution feature maps. Lin et al. [88] proposed a multi-path neural network named refinement network (RefineNet). RefineNet is an encoder decoder architecture inspired by residual connection design [34] and consists of three components; Residual convolution unit (RCU), Multi-resolution fusion and Chained residual pooling. Multi-path network exploits features at multiple levels, it refines low-resolution features with concentrated low-level features in a recursive manner to produce high-resolution feature

**Table 5**
Upsampling / Deconvolution based methods.

| Category | Strategy / Structure | | Method | Corpus | Original Architecture | Testing Benchmark | Published on | Code Available |
|---|---|---|---|---|---|---|---|---|
| Upsampling / Deconvolution | . | | Ojectness-Aware Segmentation (OA-Seg) [78] | Object Proposal Network (OPN) generate object proposals<br>Lightweight deconvolutional neural network (Light-DCNN) for upsampling | VGGNet | PASCAL VOC | October 15, 2016 | - |
| | Unpooling of Low Level Features or Score Maps | | Fully Convolutional DenseNet (FC-DenseNet) [51] | Built from a downsampling path, an upsampling path and skip connections.<br>The main goal is to exploit the feature reuses | DenseNet | CamVid Gatech | October 31, 2017 | YES |
| | | Encoder Decoder | ConvDeconvNet [79] | Convolution Network: Feature extractor<br>Deconvolution Network: Shape Generartor from the feature extractor | VGGNet | PASCAL VOC | May 18, 2015 | YES |
| | | | SegNet [80] | Obtain higher resolution by using a set of decoders one corresponding to each encoder. | VGGNet, DeconvNet | Cityscapes, KITTI, SUN RGB-D, CamVid | October 9, 2016 | YES |
| | | | Stacked Deconvolutional Network (SDN) [81] | SDN Unit: Efficient shallow deconvolutional network stack multiple SDN units one by one with dense connections. | DenseNet | PASCAL VOC CamVid, GATECH | August 16, 2017 | |
| | | | Squeeze-SegNet [82] | DFire Module: Series of concatenation of expand module of SqueezeNet. | SqueezeNet SegNet | CamVid, Cityscapes | April 13, 2018 | - |
| | | | Fully Convolutional Network (FCN) [77]<br><br>Skip Layer Architecture | Deep filter consisting (convolution, pooling, activation functions, deconvolution) layers.<br>Upsampling: end-to-end learning by backpropagation from the pixel-wise loss.<br><br>Skip (Shallow fine layer) that combines semantic information from a deep, coarse layer with the appearance information to improve segmentation.<br>FCN32s FCN16s FCN8s | Finetuning of AlexNet, VGGNet, GoogLeNet | Cityscapes, CIFAR10, KITTI, PASCAL VOC, CamVid, ADE20K, PASCAL Context, SYNTHIA, Freiburg Forest | March 8, 2015 | YES |
| | | Feature Fusion | Fully Combine Convolutional Network (FCCN) [83] | Fusing and reusing feature maps Layer by Layer | FCN-VGG | CamVid, PASCAL VOC, ADE20K | January 4, 2018 | - |
| | | | Semantic Motion Segmentation Network (SMSNet)[84] | Motion feature component: FlowNet2 architecture[85]<br>Semantic Segmentation component: AdapNet architecture<br>Fusion component: combines both the motion and semantic features | FlowNet, AdapNet | Cityscapes, KITTI | September 1, 2017 | YES |
| | | | Dense Decoder Shortcut Connections [86] | Encoder: ResNeXt architecture<br>A decoder is made up of blocks which generate semantic features maps.<br>Multi-level fusion in single-pass inference | ResNeXt | Pascal VOC, Pascal-Context, Pascal Person-Part, NYUD, CamVid | June 22, 2018 | - |
| | | | Image Cascade Network (ICNet) [87] | Proposed a cascade feature fusion (CFF) unit | Modified PSPNet | Cityscapes | April 27, 2017 | YES |
| | Reconstruction and Refinement | | Refine Network (RefineNet) [88] | Three Components<br>1. Residual convolution unit (RCU)<br>2. Multi-resolution fusion<br>3. Chained residual pooling | ResNet | Cityscapes, ADE20K, NYUDv2, SUN-RGBD, PASCAL VOC & Context | November 26, 2016 | YES |
| | | | RGB-D Multi-level Residual Feature Fusion Network (RDFNET) [89] | Multi-modal feature fusion (MMF): the fusion of features (RGB and depth)<br>Multi-level feature refinement: Refining feature | RefineNet | NYUDv2, SUN RGB-D | December 25, 2017 | YES |
| | | Encoder Decoder | Gated Feedback Refinement Network (G-FRNet) [90] | Gate Unit: Combines low-resolution features and high-resolution features to produce contextual information.<br>Refinement unit: Generate new label maps with larger spatial dimensions. | VGGNet | CamVid, PASCAL VOC, Horse-Cow Parsing | July 1, 2017 | YES |
| | | | Label Refinement Network (LRN) [91] | Predicts semantic labels at several different resolutions in a coarse-to-fine fashion. | SegNet | CamVid, SUN RGB-D, PASCAL VOC | March 1, 2017 | - |
| | | | Laplacian Pyramid Reconstruction and Refinement (LRR) [92] | Boundary mask "inset" used for localizing object boundaries.<br>LRR-32x 16x and 8x layers | ResNet | Cityscapes, PASCAL VOC | July 30, 2016 | YES |

Ref. [85] cited in the body of table 5.

maps for semantic segmentation. Islam et al. [91] proposed a refinement structure architecture called Label Refinement Network (LRN). LRN learns to predict segmentation labels at multiple levels in the network and gradually refines the results at finer scale. LRN is an encoder decoder architecture and has supervision at multiple levels (at each stage of the decoder). Zhao et al. [87] proposed image cascade network (ICNet) which utilizes semantic information in low resolution along with details from high-resolution images efficiently. The network focuses on fusion of features from multiple layers. They proposed a cascade feature fusion (CFF)

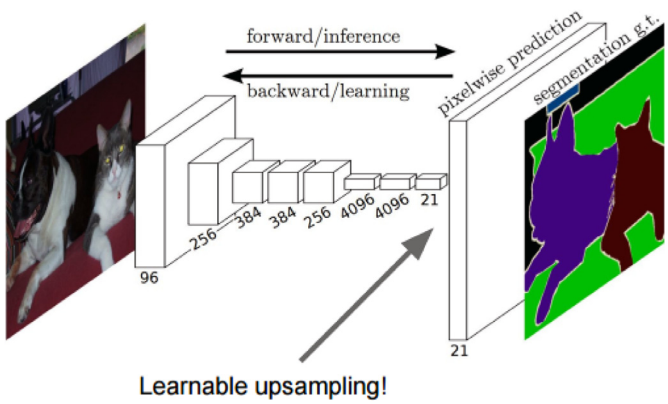



**Fig. 8.** FCN: segmentation network [77].

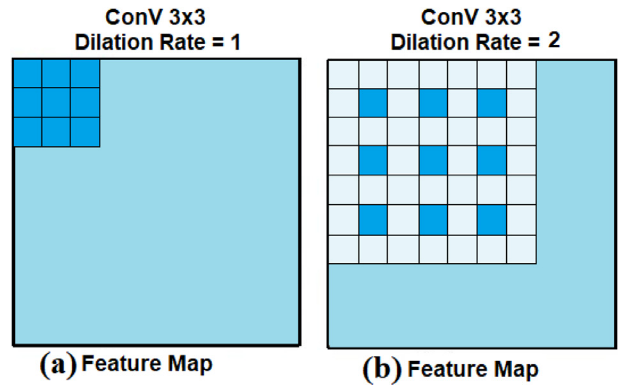



**Fig. 9.** Dilated convolution with size of $3 \times 3$ with different dilation rates. (a) dilation rate = 1, receptive field = $3 \times 3$ (b) dilation rate = 2, receptive field = $7 \times 7$.

unit that fuses the low feature maps with high feature maps. Fu et al. [81] proposed Stacked Deconvolutional Network (SDN), inspired by Huang et al. [71]. The basic idea is stacking multiple shallow deconvolutional networks one by one in order to recover high-resolution prediction. Jégou et al. [51] proposed a Fully Convolutional DenseNet FC-DenseNet, the extension of [71] by adding an upsampling path and skipping connections to recover the full input resolution. Bilinski and Prisacariu [86] proposed an architecture following encoder decoder strategy. The encoder is based on ResNeXt architecture and decoder is made of blocks (dense decoder shortcut connections), which generate semantic feature maps and allow multi level fusion in single pass inference.

Wu et al. [95] proposed a fully combined convolutional network (FCCN) to improve the upsampling operation of FCN. The network follows layer-by-layer upsampling strategy, and after each upsampling operation the size of input feature map is doubled. They also proposed a soft cost function that further improves training. Recently in [83], they extend FCCN with a highly fused network. The proposed network has three major parts: feature downsampling, combined feature upsampling and multiple predictions. The fused network makes use of multiple scale feature information in low layers. Multiple soft cost functions are used to train the proposed model. Inspired by RefineNet, Lee et al. [89] proposed RGB-D fusion network (RDFNet) for semantic segmentation. The proposed architecture is made of two feature fusion blocks: multi-modal feature fusion (MMF) to fuse features (RGB and depth) in different modalities, and multi-level feature refinement block to further refining feature for semantic segmentation. Islam et al. [90] proposed Gated Feedback Refinement Network (G-FRNet), an encoder-decoder style architecture. The proposed gated mechanism (Gate Unit) takes two feature maps one after another, i.e., low-resolution feature with larger receptive fields and high-resolution feature with smaller receptive fields, and combines them in order to produce contextual information. The feature maps with different spatial dimension generated by encoder network pass through gate unit before feeding to the decoder (feedback refinement network). The refinement network gradually refines the feature label maps. Recently, Nanfack et al. [82] proposed encoder-decoder based Squeeze-SegNet architecture. Encoder module is a SqueezeNet architecture [96] (using the fire module and removing the average pooling layer) inspired by SegNet which removes all fully connected layers of VGG. The squeeze-decoder module is the inversion of the fire module and convolutional layers of SqueezeNet.

### 2.5. Increase resolution of feature based methods

Another type of method is to recover the spatial resolution by using atrous convolution [97] and dilated convolution [98] which can generate high-resolution feature maps for dense prediction. The dilated convolution hosts another parameter “dilation rate” (describing space between the values in a kernel) to convolution layer and it has the ability to expand the receptive field without losing resolution. Table 6 shows increase resolution of feature based network models.

Chen et al. [97] from Google proposed a deep convolutional neural network model named DeepLab. Instead of using deconvolution, they proposed Atrous (‘Holes’) convolution. The atrous algorithm was originally developed by Holschneider et al. [106] for computing undecimated wavelet transform (UWT). The DeepLab architecture is similar to the one in [77] with some modification like, converting fully-connected layers into convolutional layers, using stride of 8 pixels, skip subsampling after last two pooling layers, and modifying convolutional filters in the layers (increasing length of last three convolutional layers by 2x and the first fully connected layer by 4x) by introducing zeros. The proposed method is combined with fully connected conditional random fields (CRF) and is able to produce semantically accurate predictions and detailed segmentation maps efficiently. Yu and Koltun [98], developed a convolutional neural network module design for dense prediction using dilated convolutions to combine multiscale contextual information without losing resolution and analyzing rescaled images for semantic segmentation. This module can be plugged into existing architectures at any resolution. Fig. 9 shows an example of dilation convolution with different dilation rates, which define spacing between the values in a kernel.

Treml et al. [102] proposed an encoder decoder structured architecture (SQNet). The encoder is a modified SqueezeNet architecture [96] so-called “Fire”, consisting of convolutional and pooling layers. The decoder is based on parallel dilated convolution layer. Wu et al. [105] present a fully convolutional residual network (FCRN), a new network for generating feature maps of any higher resolution, without changing the weights. They proposed a method to simulate a high resolution network with a low resolution network, and online bootstrapping method for training. In [99], Chen and his team proposed atrous spatial pyramid pooling (ASPP) module, consisting of multiple parallel atrous convolutional layers with different sampling rates to strongly segment objects at multiple scales. Fig. 10 shows example of ASPP.

The proposed network is based on the state-of-art ResNet-101 [34] image classification DCNN. They combine the network with a fully connected Conditional Random Field (CRF) in order to improve the localization of object boundaries. Yu and Koltun [104] present another deep neural network named Dilated Residual Network (DRN), a residual network ResNet [34] like architecture, in which subset of interior subsamples layers are replaced by dilation [98] to increase the resolution. The subsampling

**Table 6**
Increase resolution of features based methods.

| Category | Strategy / Structure | | Corpus | Original architecture | Testing benchmark | Published on | Code available |
|---|---|---|---|---|---|---|---|
| Increase resolution of features | Atrous convolution | DeepLab [97] | Atrous ('Holes') convolution | FCN-VGG | Cityscapes, PASCAL VOC | June 7, 2016 | Yes |
| | | DeepLabV2 [99] | Atrous Spatial Pyramid Pooling (ASPP). Method effectively enlarge the field of view of filters to incorporate multi-scale context. | FCN-ResNet | Cityscapes, PASCAL VOC, COCO | May 12, 2017 | Yes |
| | | DeepLabV3 [100] | Rethink atrous convolution augment the Atrous Spatial Pyramid Pooling (ASPP). | DeepLabV2 | Cityscapes, PASCAL VOC | December 5, 2017 | – |
| | | DeepLabV3+ [101] | Encoder Decoder Approach Xception [27] | DeepLabV3 | PASCAL VOC | March 8, 2018 | Yes |
| | Dilated convolution | Dilated Convolutions Module [98] | Rectangular Prism convolutional layers, with no pooling or subsampling for multi-scale context aggregation [34]. | VGGNet | Cityscapes, PASCAL VOC | April 30, 2016 | Yes |
| | | SQ Network [102] | Fire module: modified SqueezeNet [96] Parallel dilated convolution layer. Refinement module: SharpMask approach | SqueezeNet | Cityscapes | December 10, 2016 | – |
| | | Hybrid Dilated Convolution (HDC) [103] | Dense Upsampling Convolution (DUC) by TuSimple. | ResNet + DUC | KITTI, PASCAL VOC | November 9, 2017 | Yes |
| | | Dilated Residual Network (DRN) [104] | Replacing dilated convolutions layers into ResNet model. | ResNet | Cityscapes | May 28, 2017 | Yes |
| | Fully Convolutional Residual Network (FCRN) [105] | | Method to simulate a high resolution network with a low resolution network. Enlarge the field-of-view (FoV) of features. Online bootstrapping method for training. | ResNet + FCN DeepLab | Cityscapes, PASCAL VOC | April 15, 2016 | – |

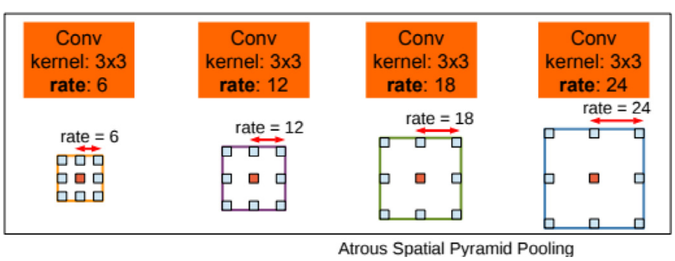



**Fig. 10.** Atrous Spatial Pyramid Pooling (ASPP) [99].

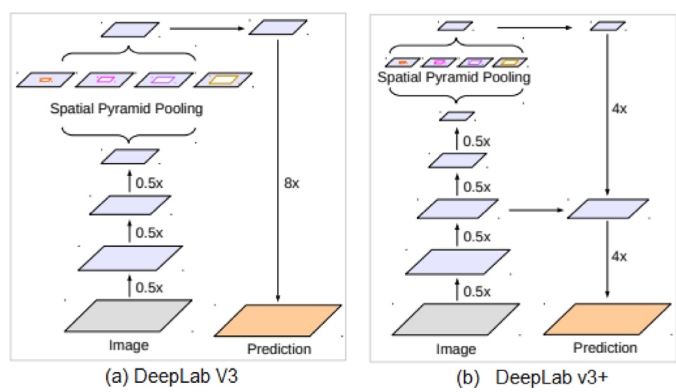



**Fig. 11.** DeepLabV3 and DeepLabV3+ [101].

removing means removing striding from some of the interior layers, increasing downstream resolution and reducing the receptive field in subsequent layers. They also propose an approach to remove the gridding artifacts introduced by dilation (degridding), which further improves the performance. Later, Chen et al. [100] revisited atrous convolution and proposed a new system network called DeepLab V3. They designed new modules in which atrous convolution works in cascade or in parallel manner (spatial pyramid pooling as shown in Fig. 11 (a)) to capture multi-scale context by adopting multiple atrous rates, and used batch normalization to train. Their main idea was to duplicate several copies of the last block in ResNet [34] and arrange them in cascade manner. Wang et al. [103] proposed a method named design dense upsampling convolution (DUC). The basic idea of DUC is to transform the label map into a smaller label map with multiple channels (dividing the label map into equal subparts having same height and width as the incoming feature map). They also proposed a hybrid dilated convolution (HDC) framework in the encoding phase that effectively enlarges the receptive fields of the network to aggregate global information. Recently in [101] the DeepLab V3+, which is the extended version of DeepLab V3 was presented. Inspired by Alvarez et al. [107], the authors proposed a decoder module, in which the encoder features are upsampled by a factor of 4 instead of 16 as in [100], then are concatenated with the corresponding low-level features from network backbone having the same spatial resolution as shown in Fig. 11 (b). They adopted the Xception model [31] and applied depth-wise separable convolution (to reduce computation complexity) to both Atrous Spatial Pyramid Pooling (ASPP) and decoder modules.

Compared to regular convolution with larger filters, atrous convolution allows to effectively enlarging the field of view of filters without increasing the number of parameters or the amount of computation. Dilated convolution is a simple yet powerful alternative to deconvolutional in dense prediction tasks.

**Table 7**
Enhancement of features based methods.

| Category | Strategy / Structure | | Corpus | Original architecture | Testing benchmark | Published on | Code available |
|---|---|---|---|---|---|---|---|
| Enhancement of features | Multi-scale features extraction | Multi-Scale Network [108,109] | Multi-scale convolutional network extract dense feature vectors that encode regions of multiple sizes centered on each pixel. Multiple post-processing methods for labeling. | LeNet | Sift Flow, Barcelona, Stanford Background | October 24, 2012 | – |
| | | | Learn multi-scale features using the image depth information. | LeNet | NYUDv2 | March 14, 2013 | – |
| | | Multi-scale Patch Aggregation (MPA) [110] | Multi-scale Patch Generator: Croppingcorresponding feature grids from Image, and aligning these grids to improve the generalization ability. A strategy is proposed to assign the classification and segmentation labels to the patches. | VGG-16 | PASCAL VOC, COCO | June 1, 2016 | – |
| | | Hypercolumns [111] | Hypercolumn classifier: Pixel classification. | Tested with R-CNN | PASCAL VOC | November 22, 2014 | – |
| | | DeepLab Attention Model [67] | Learns to weight the multi-scale features according to the object scales presented in the image, then for each scale outputs a weight map which weights feature pixel by pixel. | DeepLab | PASCAL VOC, COCO | June 1, 2016 | – |
| | | Pyramid Scene Parsing Network (PSPNet) [112] | Pyramid pooling module consists of the large kernel pooling layers for global scene prior construction | ResNet Dilated FCN | ImageNet, Cityscapes, ADE20K, PASCAL VOC | April 25, 2017 | Yes |
| | | Cascade Dilated Convolutions Network [113] | Cascading dilated convolutions (consecutive layers connection) to extract dense features. Feature fusion through Maxout Layer (Maxout Network [114]) | Dialted-ResNet FCN-VGG | PASCAL VOC | February 21, 2018 | – |
| | | Context Contrasted Local (CCL) Model [115] | CCL: Consists of several chained context-local blocks to make multi- level context contrasted local features. Gate Sum: Fusion strategy to aggregate appropriate score maps. | ResNet | Pascal Context, SUN-RGBD, COCO Stuff | June 18, 2018 | – |
| | Feature extraction from sequence of nested regions | Cascaded Feature Network (CFN) [116] | Context-aware Receptive Field (CaRF): to aggregate convolutional features of local context into strong features. | FCN + RefineNet | NYUDv2, SUN-RGBD | December 25, 2017 | – |
| | | Zoom Out [117] | Zoom out features construction using superpixels (SLIC Method) from different levels of spatial context Local Level: Superpixel itself Distant Level: Regions large enough to cover fractions of an object or entire object. Scene Level: Entire scene Combining features across levels rather than predicting. | VGG-16 | PASCAL VOC | December 2, 2014 | – |

### 2.6. Enhancement of features based methods

Enhancement of feature based methods include extraction of feature at multi-scale or from a sequence of nested regions. In deep networks for semantic segmentation, CNNs are applied to image square patches, often called kernel of fixed size centered at each pixel, labeling each pixel by observing small region around it. The network covering large and wide context (size of receptive field) is essential for better performance, which can be achieved but with increase the computational complexity. Multi-scale feature extraction or extraction from a sequence of nested region strategies can be taken in to account while ensuring computational efficiency. Table 7 shows enhancement of features based network models.

Alvarez et al. [107] propose a network algorithm to learn local features at multi-scales and multi-resolutions using different kernel sizes. The features are fused using weighted linear combination (features of each class with different weight) learned at the same time (offline) directly from the training data. Farabet et al. [108] proposed a method that extracts multiscale features vectors from the image pyramid (Laplacian pyramid version of the input image) using the multi-scale convolutional network shown in Fig. 12. Each feature vector encodes regions of multiple sizes centered on each pixel location, covering wide context.

Couprie et al. [109] adopted a similar approach, and proposed a convolutional network to learn multi-scale features using image depth information. Liu et al. [110] proposed the strategy named Multi-scale Patch Aggregation (MPA). The proposed network generates multi-scale patches for object parsing, achieves segmentation and classification for each patch at the same time and aggregates them to infer objects. Hariharan et al. [111] proposed a pixel classification method (multiple levels of abstraction and scale),

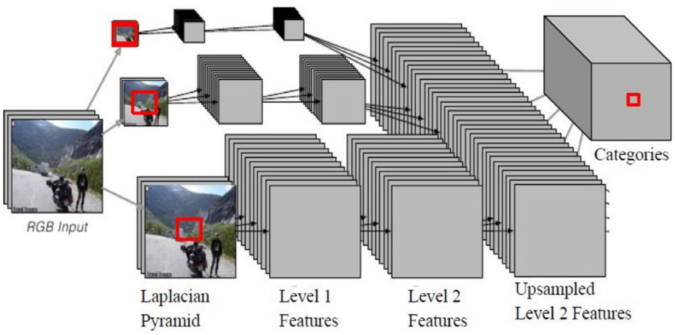



**Fig. 12.** Multiscale CNN for scene parsing [108].

Hypercolumn. The basic idea is to extract feature information from earlier layers and last layers of the CNN to allow precise localization and high semantics, and then resizing each feature map with bilinear interpolation. Further some or all of the features are concatenated into a single vector for every location.

Mostajabi et al. [117] present a feedforward classification method named Zoom-Out using Superpixels (SLIC [118]). It extracts features from different levels (local level: superpixel itself; distant level: regions large enough to cover fractions of object or entire object; scene level: entire scene) of spatial context around the superpixel to contribute to labeling decision at that superpixel. Then it computes feature representation at each level and combine all the features before feeding them to a classifier. Chen et al. [67] proposed attention based model, with ability to choose each time, which part of the input to look at in order to perform the task. The proposed attention model learns to weight the multi-scale features according to the object scales presented in the image (e.g. the model learns to put large weights on features at a coarse scale for large objects). Then for each scale, the attention model outputs a weight map which weights features pixel by pixel, and the weighted sum of FCN-produced score maps across all scales is then used for classification.

Zhao et al. [112] present pyramid scene parsing network (PSPNet) for semantic segmentation, which allows multi-scale feature ensembling. They have introduced the pyramid pooling module consisting of large kernel pooling layers shown in Fig. 12, which empirically proves to be an effective global contextual prior, containing information with different pyramid scales and varying among different sub-regions. It concatenates the feature maps with the up sampled output of parallel pooling layers. The idea is also called intermediate supervision. The representations are fed into a convolution layer to get the final per-pixel prediction. Fig. 13 shows PSPNet Architecture.

Vo and Lee [113] proposed a deep network architecture with multi-scales dilated convolution layers to extract multi scale features from multi resolution input images. The basic idea consists of cascading dilated convolutions (consecutive layers connection), each layer, with a higher rate than the previous one, achieves denser feature maps. All feature maps are then sized to same resolution and fused into a Maxout layer [114] to get most driven and leading features from all feature maps. Lin et al. [116] proposed a network called cascaded feature network (CFN). It utilizes depth information, dividing the image into layers representing visual characteristic of objects and scenes (multi-scene resolutions). Proposing context-aware receptive field CaRF (superpixel based), aggregates convolutional features of local context into strong features. The CaRF generates contextual representations, large superpixels for low scene-resolution regions and finer super pixels for regions with higher scene-resolution. Recently, Ding et al. [115] proposed a context contrasted local (CCL) model to obtain multi-scale features (both context and local). Instead of using simple sum, they proposed Gate-Sum fusion strategy to aggregate appropriate score maps, which allows a network to choose better and more desired scale of features.

Several methods aimed to capture multi-scale features, higher-layer feature contains more semantic meaning and less location information. Combining the advantages of multi-resolution images and multi-scale feature descriptors to extract both global and local information in an image without losing resolution improves the performance of the network.

### 2.7. Semi and weakly supervised concept

The CNN's are becoming deeper and deeper by increasing the depth and width (the number of levels of the network and the number of units at each level). Deep CNN requires large-scale dataset and massive computing power for training. Collecting labeled dataset manually is time consuming and requires enormous human efforts. To comfort these efforts, semi or weakly supervised methods are applied using deep learning techniques. Table 8 shows semi and weakly supervised network models used for semantic segmentation.

Work by Pathak et al. [119] is to be the first considering the fine-tuning of CNN pre-trained for object recognition, using image-level labels, within a weakly supervised segmentation context. They introduced a fully convolutional network method, which relies on a Multiple Instance Learning (MIL-FCN) [138], i.e., learn pixel-level semantic segmentation from weak image level labels indicating the presence or absence of an object. They proposed a multi-class pixel-level loss inspired by the binary MIL scenario. Pinheiro and Collobert [120] proposed a weakly supervised approach to produce pixel level labels from image-level labels using Log-Sum-Exp (LSE) [121] method, which assigns the same weight to all pixels of the image during the training. Papandreou et al. [123] presented a weakly and semi-supervised learning method using weak annotations, either alone or in combination with small number of strong annotations. They developed a method called Expectation Maximization (EM) for training DCNN from weakly annotated data. Hong et al. [122] proposed a semi-supervised method (DecoupledNet), which uses two separate networks, one for classification (classifies the object label) and the other for segmentation (to obtain figure-ground segmentation of each classified label). Dai et al. [135] propose a method based on bounding box annotations (BoxSup). The unsupervised region proposal method (selective search [53]) is used to generate segmentation masks, and these masks are used for training convolutional network. The proposed BoxSup model, trained with a large set of boxes, increases the object recognition accuracy (the accuracy in the middle of an object), and improves object boundaries. Khoreva et al. [139] proposed a box-driven segmentation technique for semantic segmentation, which generates input labels for training from the bounding box annotations using Grab Cut-like algorithm [140] without modifying the training procedure. Luo et al. [125] present a weakly and semi-supervised dual image segmentation (DIS) learning strategy, which performs segmentation (capturing the accurate object classes), and reconstruction (accurate object shapes and boundaries). The idea is to predict tags, label maps from an input image, and perform reconstruction of images using predicted label maps.

Saleh et al. [129] proposed weakly supervised segmentation network with built-in foreground/background prior. The main idea is to extract the localization information directly from the network itself (extracting foreground/background masks). Later in [130], they extended their work to obtain multi-class (class-specific) masks by the fusion of foreground / background ones with information extracted from a weakly supervised localization network inspired by Zhou et al. [141]. Saito et al. [131] present a method

**Table 8**
Semi and weakly supervised based methods.

| Category | Strategy/ Structure | | | Corpus | Original architecture | Testing benchmark | Published on | Code available |
|---|---|---|---|---|---|---|---|---|
| Weakly and semi supervised | Image level labels | Multiple Instance Learning (MIL-FCN) [119] | | Multi-class pixel-level loss inspired by the binary MIL scenario. | VGG | PASCAL VOC | April 15, 2015 | – |
| | | Aggreg-LSE [120] | | An approach to produce pixel-level labels from image-level labels using Log-Sum-Exp (LSE) [121]. | VGG | PASCAL VOC | June 7, 2015 | – |
| | | Utilization of heterogeneous annotations | DecoupledNet [122] | Classification network: Identifies labels Segmentation network: Produces pixel-wise figure-ground segmentation corresponding to each identified label. Bridging layers connecting the two Networks (Decoupling). | VGG | PASCAL VOC | June 17, 2015 | Yes |
| | | | WSSL [123] | Expectation Maximum (EM) module for fast training under both weakly and semi-supervised settings. | DeepLab | Cityscapes, PASCAL VOC | December 7, 2015 | Yes |
| | | Simple to Complex (STC) [124] | | A progressively training strategy is proposed by incorporating simple-to-complex images with image-level labels. | VGG + DeepLab | PASCAL VOC | November 1, 2017 | – |
| | | Dual Image Segmentation DIS [125] | | Segmentation: Predict tags and label maps from the image (captured the accurate object classes). Reconstruction: The reconstruction of images using predicted label maps (accurate object shapes and boundaries). | ResNet | PASCAL VOC | December 25, 2017 | – |
| | | Adversarial learning | SW-GAN [126] | Generative Adversarial Network framework which extends the typical GAN to a pixel-level prediction. | VGG | PASCAL VOC, SiftFlow, StanfordBG, CamVid | March 28, 2017 | – |
| | | | Semi-Adv [127] | Propose a fully convolutional discriminator that learns to differentiate between ground truth label maps and probability maps of segmentation predictions. | DeeplabV2 | PASCAL VOC, Cityscapes | February 22, 2018 | Yes |
| | | Segmenting Path Proposals [128] | | Weakly-supervised approach to segmenting proposed paths for a road vehicle Method for generating a large amount of labeled images without any manual annotation. | SegNet | KITTI, Oxford | November 17, 2017 | – |
| | | Built-in feature extraction approach | Fg/Bg Masks [129] | Weakly-supervised segmentation network with built-in Foreground/Background Prior “Information extracted from a pre-trained network”. | VGG-16 | PASCAL VOC | September 2, 2016 | – |
| | | | Multi-Class Mask [130] | Foreground/background mask combined to generate the class-specific mask Multi-Class Prior. | VGG-16 | PASCAL VOC | June 6, 2017 | – |
| | | | Superpixel Clustering Method [131] | Pre-trained Dilated ResNet for Feature extraction SuperPixel Align Method (FH Superpixel) Road Feature Clustering (K-Means). | DRN + SegNet | Cityscapes | November 16, 2017 | – |
| | | Deep Seeded Region Growing (DSRG) Network [132] | | Utilize the Seeded Region Growing mechanism to generates pixel-level labels. | VGG | PASCAL VOC, MS COCO | February 1, 2018 | Yes |
| | | Multi-Dilated Convolutional (MDC) [133] | | Multi-Dilated Convolutional (MDC) Blocks: Produce dense object localization maps which can be utilized for segmentation both in weakly and semi-supervised manner. | VGG + DeepLab | PASCAL VOC | May 28, 2018 | – |
| | Multi-level labels | Diverse Supervision | Annotation-Specific FCN [134] | Annotation-Specific Loss Module Image-level labels for classification Box-level labels for object detection Pixel-level labels for semantic segmentation | FCN | PASCAL VOC | February 1, 2018 | – |
| | Bounding box | Boxsup [135] | | The semi-supervised approach based on bounding box annotations Uses SelectiveSearch [136]: to generate segmentation masks. Iterate between an automatically generating region proposals and training convolutional network | FCN | PASCAL VOC, CONTEXT, MS COCO | May 17, 2015 | – |
| | | MCG-GrabCut+ [137] | | A weakly supervised approach based on bounding box annotations Uses GrabCut+ Approach [132]: to estimate object segment. | VGG + DeepLab | PASCAL VOC, MS COOC | November 9, 2017 | Yes |

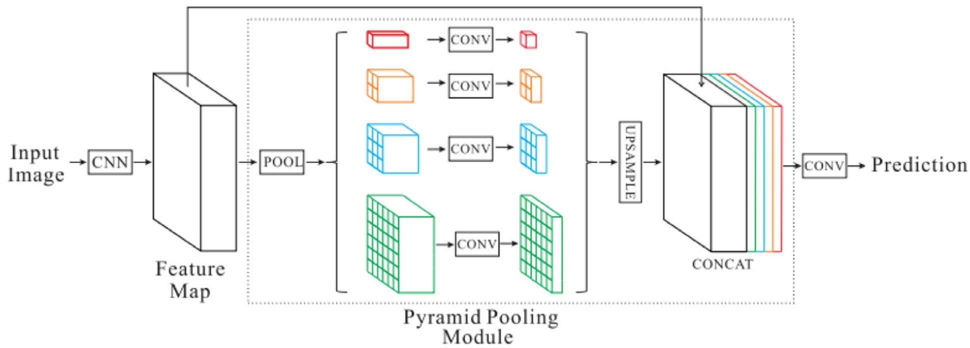



**Fig. 13.** Pyramid Scene Parsing Network (PSPNet) [112].

that uses the feature maps extracted from a pre-trained dilated ResNet having built-in priors for semantic segmentation. They proposed a superpixel clustering method to generate road clusters (to select largest cluster at the bottom half of image), that are considered as the label to train CNN for segmentation. Barnes et al. [128] develop a weakly supervised method for autonomous driving applications for generating a large amount of labelled images (from multiple sensors and data collected during driving) containing path proposals without any manual annotation. Ye et al. [134] proposed a method for learning convolutional neural network models from images with three different types of annotations, i.e., image-level labels for classification, box-level labels for object detection and pixel-level labels for semantic segmentation. They proposed an annotation-specific loss module (with three branches, each branch with a different loss function), which is designed to train the network for each of the three different annotations.

Souly et al. [126] proposed a semi-supervised semantic segmentation method using adversarial learning inspired by Generative Adversarial Networks (GANs) [142]. Later, Yurdakul and Yemez [154] proposed a similar approach which consists of two sub nets; segmentation net (to generate class probability maps) and discriminator net (to generate spatial probability maps with both labeled and unlabeled data). Wei et al. [133] presented a weakly and semi supervised approach by using multiple dilated convolutions. They proposed augmented classification network with multi-dilated convolutional (MDC) blocks that generate dense object localization maps, which are utilized for semantic segmentation in both weakly and semi supervised manner. Huang et al. [124] proposed a weakly supervised network, which produces labels using the contextual information within an image. They proposed a seeded region growing module to find small and tiny discriminative regions from the object of interest using image labels to generate complete and precise pixel level labels, which are used to train semantic segmentation network. Wei et al. [143] proposed a Simple to Complex (STC) network, a weakly supervised approach using image-level annotations. The basic idea is first to learn from simple images (to generate saliency maps using discriminative regional feature integration (DRFI)), and then apply learned network to the complex images (to generate pixel-level segmentation masks of complex images) for semantic segmentation.

Semi and weakly supervised learning aims to reduce the load for full annotation. These methods improved learning performance using weak annotations in the form of image-level labels (information about which object classes are present) and bounding boxes (coarse object locations).

### *2.8. Spatio-temporal based methods*

In this section, we aim to investigate the deep convolutional networks that use spatial information along with temporal information for semantic segmentation.

In a video, frames are associated with each other and have temporal information (i.e., features of continuous sequences of frames) that can be useful for interpreting a video semantically. Spatio-temporal structured prediction can prove useful in both supervised and semi-supervised manner. Table 9 shows Spatio-Temporal based network models for semantic segmentation.

Several methods are proposed in the combination of Recurrent Neural Networks (RNN) and Convolutional Neural Network (CNN) for video segmentation. Fayyaz et al. [145] presented a full convolutional network Spatio-Temporal Fully Convolutional Network (STFCN) employing spatial and temporal features. They proposed spatio-temporal module that takes the advantage of LSTM in order to define temporal features. The spatial feature maps of the region in single frame fed into LSTM, infers a relation with spatial features of equivalent regions in frames before that frame. Further, spatial and temporal information fed into dilated convolution network ([98] with minor modifications) for upsampling and are fused (summing operation) for semantic predictions. He et al. [146] proposed Spatio-temporal data-driven pooling model (STD2P), which is method to integrate multi-view information by using super pixels and optical flow. The goal of multi-view semantic segmentation is to make use of the potentially richer information from different views with better segmentations than single view. Qiu et al. [148] proposed 2D/3D FCNs based architectural model named deep spatio-temporal ful-ly convolutional networks (DST-FCN), that utilizes spatial and temporal dependencies among pixels and voxels. The proposed architecture is a two stream network, Sequential frame stream, (2DFCN for spatial and ConvLSTM for temporal information), and clip stream, (3DFCN based on C3D [152] developed on voxel level). Pavel et al. [153] present a recurrent convolutional neural network model utilizing spatial and temporal information for processing image sequences. Yurdakul and Yemez [154] proposed a network that combines color and depth information in RGBD videos for semantic segmentation using convolutional and recurrent neural network frameworks.

Some architectures are based on Gated Recurrent Architectures to overcome gradients problem. Ballas et al. [155] used a term percepts (visual representations extracted from different levels of DCN) to capture spatial-temporal feature information in the video using gated-recurrent-unit recurrent networks. Siam et al. [149] present a fully convolutional network based on gated recurrent architecture (RFCN). Three different architectures were used following two approaches, conventional recurrent units (RFCLeNet) and convolutional recurrent units (RFC VGG, RFC Dilated), learning spatio-temporal features with less number of parameters. Nilsson et al. [151] present Gated Recurrent Flow Propagation network. They proposed Spatio Temporal Transformer Gated Recurrent Unit (STGRU), combining the strength of spatial transformer (for optical flow warping) with convolutional gated architecture (to adaptively propagate and fuse estimates). Shelhamer et al. [144] proposed a network named Clockworks, which

**Table 9**
Spatio-temporal based methods.

| Category | Strategy / Structure | Corpus | Original architecture | Testing benchmark | Published on | Code available |
|---|---|---|---|---|---|---|
| Spatio-temporal | Clockwork FCN [144] | Clockworks: clock signals that control the learning of different layers with different rates | FCN Clockwork RN | Youtube-Objects, NYUD, Cityscapes | August 11, 2016 | Yes |
| | Spatio-Temporal FCN [145] | Spatial-temporal module embedding into FCN LSTM to define relationships between image frames | FCN | Camvid NYUDv2 | September 2, 2016 | Yes |
| | Spatio-Temporal Data-Driven Pooling (STD2P) [146] | Incorporate superpixels and multi-view information into convolutional networks | FCN | NYUDv2 SUN 3D | April 26, 2017 | – |
| | Feature Space Optimization (FSO) [147] | Optimize the mapping of pixels to a Euclidean feature space used by DenseCRF for spatio-temporal regularization | VGG Dilation | CityScapes, Camvid | December 12, 2016 | Yes |
| | Deep Spatio-Temporal FCN (DST-FCN) [148] | Learn spatial-temporal dependencies through 2D FCN on pixels and 3D FCN on voxels | VGG C3D | A2D, CamVid | October 5, 2017 | – |
| | Gated Recurrent FCN [149] | Implementation of three gated recurrent architectures RFC-LeNet: Conventional recurrent units. RFC-VGG and RFC-Dilated: Convolutional recurrent units. | FCN | SegTrack V2, Davis, Cityscapes, SYNTHIA | November 21, 2016 | – |
| | WSBF [150] | Weakly-supervised two-stream network. One stream takes image, and other optical flow to extract the features. RFC-VGG and RFC-Dilated: Convolutional recurrent units. | VGG | Cityscapes, CamVid, YouTube-Objects | August 15, 2017 | – |
| | Gated Recurrent Flow Propagation (GRFP) [151] | Spatio-Temporal Transformer Gated Recurrent Unit (STGRU) Combining spatial transformer with convolutional-gated architecture. | Dilation LRR | CityScapes, Camvid | October 2, 2017 | – |

is a combination of FCN and clockwork recurrent network [156], grouping the layers of the network into stages with different rates (either fixed clock rate or adaptive clock) and then fusing them via skip connections. Saleh et al. [150] proposed a weakly supervised framework for video semantic segmentation that treats both foreground and background classes equally. The basic idea is to manage multiple foreground objects and multiple background objects equally. They propose an approach to extract class-specific heatmaps from classifier that localizes the different classes for both without pixel level or bounding box annotations. Kundu et al. [147] proposed a model to optimize the feature space used by the fully connected conditional random field for spatio-temporal regularization. Chandra et al. [157] proposed a Video Gaussian Conditional Random Field approach for spatio-temporal structured prediction, which is an extension of [158]. The FCN network obtains unary (class score per-pixel), spatial pairwise and temporal pairwise terms, which are fed into G-CRF module that performs inference (linear system) to obtain the final prediction.

### 2.9. Methods using CRF / MRF

Semantic segmentation involves pixelwise classification and such pixelwise classification often produces unsatisfactory results (poor, incorrect and noisy predictions) that are irreconcilable with the actual visual features of the image [159].

Markov random field (MRF) and its variant Conditional Random Fields are classical frameworks that are widely used to overcome these issues. They express both unary term (per-pixel confidence of assigning labels) and pairwise terms (constraints between adjacent pixels). CNNs can be trained to model unary and pairwise potentials in order to capture contextual information. The context provides important information for scene understanding tasks such as spatial context which provides semantic compatibility/incompatibility relation between objects, scenes and situations. CRFs can be a post processing or end-to-end, to smooth and refine the pixel prediction in semantic segmentation. They combine class scores from classifiers with the information captured by the local interactions of pixels and edges or superpixels. Table 10 shows network models using CRF.

Krahenbuhl et al. [160] proposed a fully connected CRF (DenseCRF) model, in which pairwise edge potentials are defined by a linear combination of Gaussian kernels. The method is based on mean field approximation, message passing is performed using Gaussian filtering techniques [161]. Methods [79,97,123,129,133,135,139,150] coupled fully connected CRF with their proposed DCNNs to produce accurate predictions and detailed segmentation maps for improving performance. Zheng et al. [162] formulate mean-field inference algorithm for the dense CRF with Gaussian filtering technique as recurrent neural network (CRF-RNN), that performs CRF-based probabilistic graphical modeling for structured prediction. Fig. 14 shows CRF as RNN.

Vemulapalli et al. [163] proposed a model named Gaussian Mean Field (GMF) network that models unary potentials, pairwise potentials and Gaussian CRF inference for the task of semantic segmentation. In the proposed network, output of each of the layer is closer to maximum a posteriori probability (MAP) estimated to its input. Chandra et al. [158] proposed a Gaussian Conditional Random Field (G-CRF) module using a quadratic energy function that captures unary and pairwise interactions. Lin et al. [169] propose a model Context CNN CRF jointly learning CNN and CRFs. They formulate CRF with CNN pairwise potential to capture contextual relationship between neighboring patches and sliding pyramid pooling (multi-scale image network input) for capturing patch-background context that can be combined to improve the segmentation. Instead of learning the potentials, [168] proposes a method that learns CNN message estimators for the message passing inference for structured Conditional Random Field (CRFs) predictions. Teichmann and Cipolla [164] proposed

**Table 10**
Methods using CRF/MRF.

| Category | Strategy / Structure | Corpus | Original architecture | Testing benchmark | Published on | Code available |
|---|---|---|---|---|---|---|
| CRFs / MRFs | Fully Connected-CRF (DenseCRF) [160] | Based on mean field approximation, message passing performed using Gaussian filtering techniques [161]. | ResNet | PASCAL VOC | May 15, 2018 | Yes |
| | CRF-RNN [162] | Multiple mean-field iterations. Interpretation of dense CRFs as Recurrent Neural Networks (CRF-RNN) combined with CNN. | FCN | PASCAL VOC Cityscapes | April 13, 2016 | – |
| | Gaussian Conditional Random Field (GCRF) / Gaussian Mean Field (GMF) Network [163] | GMF Network: Performing Gaussian mean field inference. | DeepLab | PASCAL VOC ImageNet | June 26, 2016 | Yes |
| | Quadratic Optimization (QO) [158] | Quadratic Optimization (QO) module | FCN | PASCAL VOC | November 29, 2016 | – |
| | Convolutional-CRF (ConvCRF) [164] | Inference in terms of convolutions. | ResNet | PASCAL VOC | May 15, 2018 | Yes |
| | Incorporating higher order potentials / Higher-order CRF [165] | Object-detection based potentials: Provide semantic cues for segmentation. Superpixel-based potentials: Encourage label consistency over regions. | CRF-RNN | PASCAL VOC, Context | July 29, 2016 | – |
| | Structured Patch Prediction (SegModel) [166] | Integrate segmentation specified features, high order context and boundary guidance. | FCN | PASCAL VOC Cityscapes ADE20K | November 9, 2017 | – |
| | Deep Parsing Network (DPN) [167] | Models Unary term and Pairwise terms in single CNN. | VGG | PASCAL VOC | September 24, 2015 | – |
| | Learning Messages [168] | CNN message estimators for the message passing inference. | VGG-16 | PASCAL VOC | September 8, 2015 | – |
| | Adelaide / Bounding-box detection / Adelaide Very Deep FCN [136] | Hough transform based approach Online bootstrapping method for training. | FCRN | PASCAL VOC | May 23, 2016 | – |
| | Context CNN CRF [169] | Patch-patch context: Formulate CRFs to capture contextual relationship between neighboring patches Patch-background context: Sliding Pyramid Pooling. | VGG-16 | PASCAL VOC NYUDv2 Pascal Context Siftflow | June 6, 2016 | – |
| | Incorporate the depth information / Depth-sensitive fully-connected Conditional Random Field (DFCN-DCRF [170]. | Fully-connected CRFs with RGB information and depth information. | FCN | SUN-RGBD | October 4, 2017 | – |

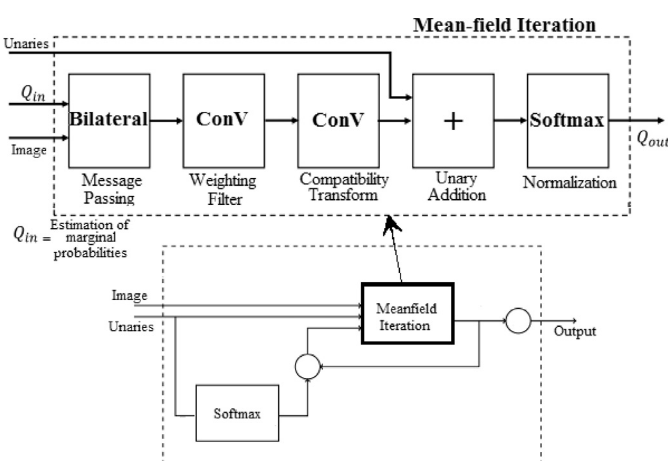



**Fig. 14.** CRF as a recurrent neural network [162].

convolutional CRFs (ConvCRFs) method that reformulates the message passing inference in terms of convolutions.

Some methods employed higher-order potentials (based on object detection or superpixels) modeled as CNN layers when using mean field inference and effective in improving semantic segmentation performance. Arnab et al. [165] proposed a method in which CRF models unary and pairwise potentials together with high-order potentials object detector (to provide semantic cues for segmentation) and superpixel (having label consistency over regions) in an end-to-end trainable CNN. Shen et al. [166] proposed joint FCN and CRF model (SegModel) that integrates segmentation specified features, which constitutes high order context and boundary guidance (bilateral-filtering based CRF) for semantic segmentation. Liu et al. [167] proposed Deep Parsing Network (DPN), which models unary term and pairwise terms (i.e., high-order relations and mixture of label contexts) in single CNN that achieve high performance by extending the VGG network, and adding some layers for modeling pairwise terms. Jiang et al. [170] utilize the depth information as complementary information into conditional random fields. They proposed depth sensitive fully connected conditional random field combined with a fully convolutional network, (DFCN-DCRF). The basic idea is to add the depth information into dilated-FCN and fully connected CRF to improve accuracy for semantic segmentation.

CRF inference with deep convolutional neural network improves pixel-level label predictions by producing sharp boundaries and dense segmentation. Several methods learn arbitrary potentials in CRFs. It has been used as post processing, end-to-end fashion, formulated as RNN and incorporated as module in existing neural network.

### *2.10. Alternative to CRF*

Integrating conditional random field into original architecture is a difficult task due to additional parameters and highly computational complexity at training. Moreover, the majority of CRFs uses hand constructed color-based affinities that may lead to spatial false predictions. Several methods have been proposed to

**Table 11**
Alternative to CRF based methods.

| Category | Strategy / Structure | Corpus | Original architecture | Testing benchmark | Published on | Code available |
|---|---|---|---|---|---|---|
| Alternative to CRF approaches | Bilateral Neural Network (BNN) [171] | Bilateral filter inference in DenseCRF replacing Gaussian potentials with bilateral convolution to learn pairwise potentials. | DeepLab | Pascal VOC | June 26, 2016 | Yes |
| | Fast Bilateral Solver (BS) [172] | Edge-aware smoothness algorithm using bilateral filtering technique. | CRF-RNN | Pascal VOC MS COCO | July 22, 2016 | – |
| | Boundary Neural Field (BNF) [173] | Build unary and pairwise potentials from input RGB image, then combine them in global manner. | FCN | Semantic Boundaries Dataset | May 24, 2016 | – |
| | DT-EdgeNet [174] | Domain transform (DT) Module: Edge-preserving filter. Edge Net: Predicts edge features from midway layers. | DeepLab | Pascal VOC | December 12, 2016 | – |
| | Global Convolutional Network (GCN) [175] | Large kernels used for classification and localization. Boundary Refinement Block: Model the boundary alignment as a residual structure. | FCN ResNet | Cityscapes COCO PASCAL VOC | March 8, 2017 | – |
| | Random Walk Network (RWN) [176] | Random walk network pixel labeling framework | DeepLab-largeFOV | Pascal, SBD-Stanford Background, Sift Flow | July 22, 2017 | – |

overcome these issues and can be used as alternate to CRFs. Table 11 shows network models alternate to CRFs.

Bertasius et al. [173] proposed a FCN architecture named Boundary Neural Field (BNF) to predict semantic boundaries and produce semantic segmentation maps using global optimization. The BNF combines the unary potentials (prediction by FCN) and pairwise potentials (boundary-based pixel affinities) from the input RGB image in a global manner. The basic idea is to assign pixels to the foreground and background labels for each of the different object classes and apply constraint relaxation. Later in [176], they proposed Convolutional Random Walk Network (RWN) addressing same issue, model based on random walk method [177]. The network model predicts semantic segmentation potentials and pixel level affinities, and combines them through proposed random walk layer that applies spatial smoothing predictions.

Jampani et al. [171] propose a network based on Gaussian bilateral filter [178], named bilateral neural network (BNN). Bilateral filter inference in fully connected CRF [160] (by replacing Gaussian potentials with bilateral convolution) to learn pairwise potentials of fully connected CRF. Barron and Poole [172] propose edge-aware smoothness algorithm using bilateral filtering technique name the bilateral solver. Peng et al. [175] proposed a residual-based boundary refinement model, Global Convolutional network (GCN), for semantic segmentation. They proposed boundary refinement block (FCN structure without fully connected and global pooling layers) to model the boundary alignment as a residual structure. Chen et al. [174] proposed a model with domain transform (DT) module as a substitute to CRF, an edge preserving filtering method. The model consists of three modules. The first module produces semantic segmentation score prediction based on DeepLab. The second module named Edge Net, predicts edge features from midway layers and the third module is an edge-preserving filter named Domain Transform (recursive filtering), proposed in [179].

Several methods have been proposed that can be used as alternative to CRF with the advantage of fast and less number of parameters. Bilateral filtering techniques can be useful tool in the construction of deep learning frameworks.

Fig. 15 gives an overview to the readers to have good understanding of the categorization of different methods for semantic segmentation.

## 3. Datasets and evaluation for deep learning techniques

One of the hardest problem for any segmentation systems based on deep learning techniques is the collection of data in order to construct suitable dataset. There are four possible ways to get labeled data as shown in Fig. 16. Traditional Supervision: hand label data; Weak supervision: obtained automatically without human annotators using unlabeled data; Semi-supervised learning: partially labeled and partially unlabeled data, and transfer learning: using pre-trained model as a start point.

### *3.1. Datasets*

The dataset acts as the benchmark against which deep learning networks are trained and tested. Several datasets has been constructed over the last few years that are used in deep learning, motivating researchers to create new models and strategies with better generalization abilities.

These datasets can be categorized according to the nature of data.

*The automotive datasets includes;* CamVid dataset [180] which is considered as the first with semantically annotated videos, Daimler Urban Segmentation [181], CityScapes [182], Mapillary Vistas [183] and the most recent Apolloscape-Scene parsing [184] which focuses on semantic understanding of urban street scenes. The KITTI [185] dataset used in various computer vision tasks such as 2D/3D object detection, stereo, optical flow, and tracking. Synthetic datasets [186,187] consist of a thousand images extracted from realistic open-world games.

*Data sets generic in nature;* PASCAL VOC [188] is one of the most popular and widely used dataset in deep learning semantic segmentation, CIFAR-10/100 [189] contains up to 60,000 images, offering 10 and 100 categories of tiny $32 \times 32$ images. A remarkable ImageNet [190] dataset contains over 14 million labeled images, SegTrack v2 [191] is a video segmentation dataset with annotations on multiple objects at each frame, and PASCAL Context [192] is a set of additional annotations for PASCAL VOC. Microsoft-COCO [193] is a collection of images of complex everyday scenes contains common natural objects, ADE20K [194] containing both indoor and outdoor images with large variations, and DAVIS [195] dataset containing densely annotated videos with pixel accurate ground truth. Recently developed COCO stuff [196] dataset augments the original COCO dataset with much more comprehensive stuff annotations.

*Indoor environment datasets;* NYUDv2 [197] is composed of RGB-D images and video sequences from a variety of indoor scenes, Cornell RGB-D [198] contains labeled office and home scene point clouds, ScanNet [199] comprises more than 1500 scenes annotated with 3D camera pose, surface reconstructions, and semantic segmentations. Stanford 2D-3D [200] contains mutually registered

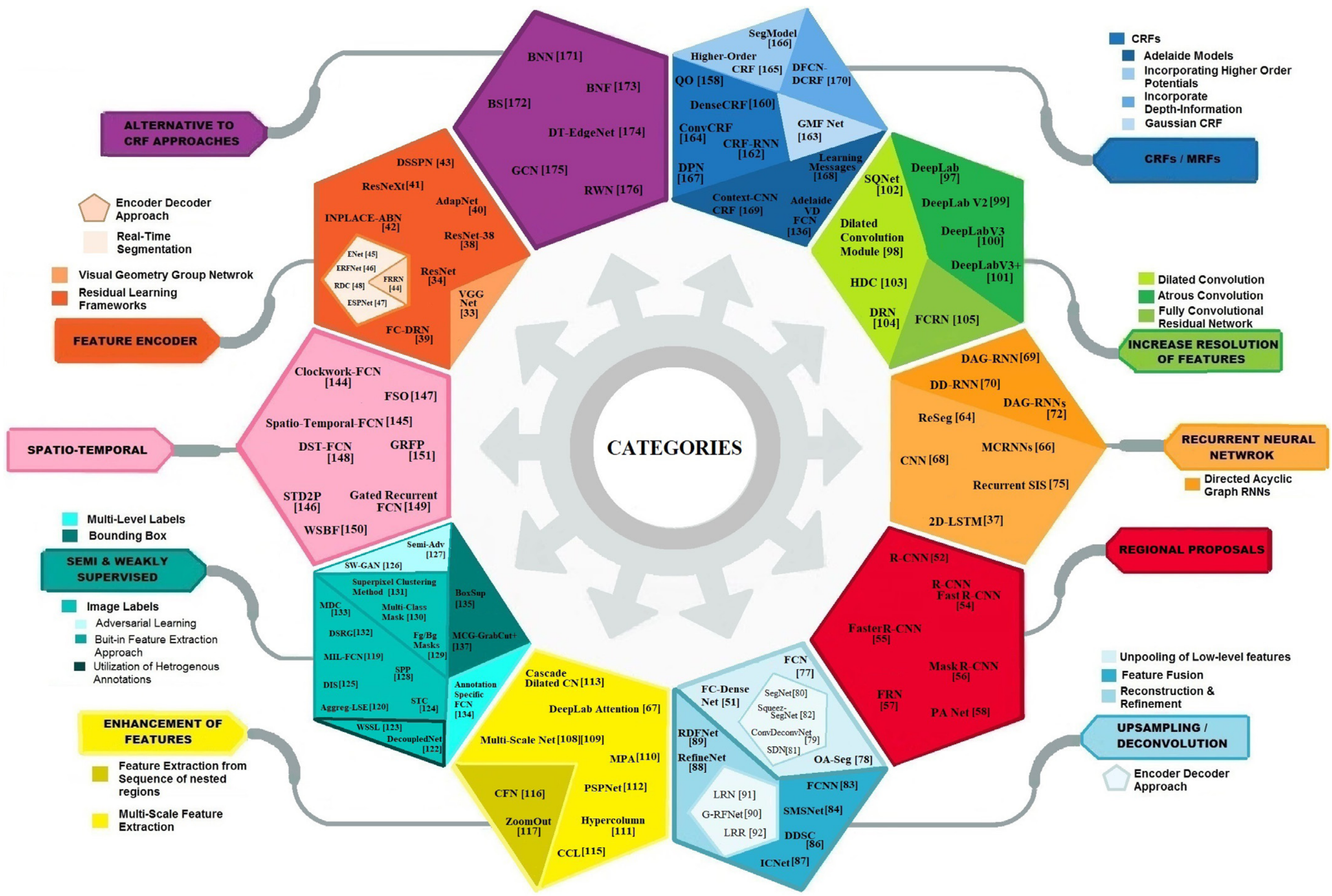



**Fig. 15.** Illustration of the ten categories into which we have classified the reviewed semantic segmentation methods.

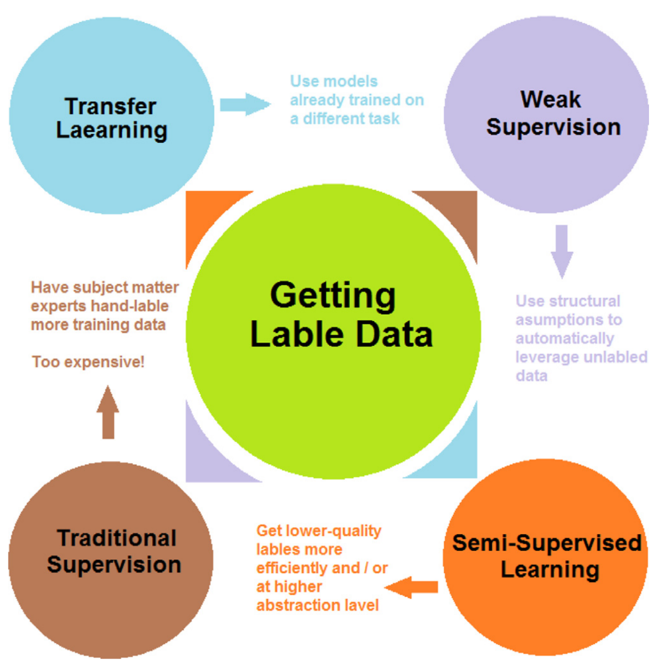



**Fig. 16.** Getting label data.

modalities from 2D/3D domains, with 71,882 RGB images (both regular and 360°), along with the corresponding depths, surface normal and semantic annotations. SUN 3D [201] and SUN RGB-D [202] datasets contain videos of big spaces for place-centric scene understanding.

*Object datasets;* RGB-D Object v2 [203] containing 25000 images of common household objects in 51 categories, YouTube Dataset [204] comprises 126 videos.

*Datasets for outdoor environment;* Microsoft Cambridge [205] consists of 591 real outdoor scene photographs of 21 object classes; Graz-02 [206] is a natural-scene object category dataset created at INRIA. LabelMe [207] contains outdoor images of 8 different classes that are taken in different cities of Spain; Barcelona dataset [208] is a subset of LabelMe; Stanford-background [209] and PASCAL SBD [210] are collected from PASCAL VOC; Sift-flow [211] consists of 2688 images of 256 × 256 pixels and 33 classes, and Freiburg Forest [212] constitute on outdoor forest environment in different condition lighting, shadows and sun angles.

The dataset construction is both time consuming and labor intensive, so for the researchers and developers the most practical and workable approach is to use existing standard datasets which are representative enough for the domain of the problem. Some datasets have become standard and commonly used by researchers to compare their work with others using standard metric for evaluation. Dataset selection at a start of research is challenging task, therefore the comprehensive description on dataset can help. In Table 12, we list the datasets used by deep learning networks that are publicly available. Are given different information such as environment nature, the number of classes, training/testing samples, image resolution, year of construction, and best performances achieved till date (to the best of our knowledge) by the models for

**Table 12**
Summary of datasets.

| Dataset | Environment nature | No of classes | Training | Samples Validation | Test | Image Resolution | Year | Performance | Network model |
|---|---|---|---|---|---|---|---|---|---|
| ADE20K [194] | Generic | 150 | 20210 | 2000 | – | Variable | 2016 | 44.98% MIoU | PSPNet [112] |
| Apolloscape Scene parsing [184] | Street view / 2D-3D | 25 | | 146997 Frames | | $3384 \times 2710$ | 2018 | | |
| Barcelona [208] | Outdoor | 170 | 14871 | – | 279 | $640 \times 480$ | 2010 | 74.6% GL acc. | DAG-RNN [69] |
| CamVid [180] | Street view | 32 | | 701 | | $960 \times 720$ | 2009 | 69.94% MIoU | FCCN [83] |
| CIFAR-10/100 [189] | Generic / Objects | 10/100 | 50K/500 | – | 10K/100 | $32 \times 32$ | 2009 | 3.58% test error | ResNeXt [41] |
| Cityscapes [200] | | | | | | | | | |
| Fine | Street view | 30 | 2975 | 500 | 1525 | $2048 \times 1024$ | 2016 | 79.3% MIoU | DeepLabV3 [100] |
| Coarse | | | 22973 | 500 | – | | | 82.2% MIoU | DeepLabV3+ [101] |
| Cornell RGB-D [198] | Indoor office/Home | – | | 24 Office / 28 home scenes point clouds | | Variable | 2011 | | |
| COCO Stuff [196] | Generic | 172 | | 163957 | | Variable | 2018 | 38.9% MIoU | DSSPN [43] |
| DAVIS [195] | Generic / Videos | 4 | 4219 | 2023 | 2180 | 480p | 2017 | 69.84% MIoU | RFCNet [149] |
| Data from Game [187] | Synthetic / Street view | 19 | | 24966 | | $1914 \times 1052$ | 2016 | | |
| Daimler Urban Segmentation [181] | Street view / Video | 5 | | 500 | | $1024 \times 440$ | 2013 | 77.2% MIoU | Layered Interpretation [213] |
| Freiburg Forest [212] | Outdoor / Forest-environment | 6 | 230 | – | 136 | $1024 \times 768$ | 2016 | 88.25% MIoU | AdapNet [40] |
| ImageNet [190] | Generic | 1 K | | 14,197,122 | | Variable | 2010 | | |
| INRIA-Graz-02 [206] | Outdoor /Natural | 3 | 479 | – | 479 | $640 \times 480$ | 2007 | | |
| KITTI [185] | Street view | 10 | 140 | – | 112 | $1226 \times 370$ | 2015 | 63.51% MIoU | LSDN [214] |
| LabelMe [207] | Outdoor | 8 | 2920 | – | 1133 | Variable | 2008 | | |
| Mapillary Vistas [183] | Street view | 66 | 18000 | 2000 | 5000 | $1920 \times 1080$ | 2017 | 45.01% MIoU | DSSPN [43] |
| Microsoft COCO [193] | Generic | 80 | 82783 | 40504 | 81434 | Variable | 2014 | 56.9% AP | FPN [57] |
| Microsoft Cambridge [13] | Outdoor | 21 | | 591 | | $320 \times 240$ | 2005 | | |
| NYUDv2 [197] | Indoor | 40 | 795 | 654 | – | $480 \times 640$ | 2012 | 50.1% MIoU | RDFNet [89] |
| PASCAL | | | | | | | | | |
| VOC [188] | Generic | 20 | 1464 | – | 1449 | Variable | 2012 | 89.0% MIoU | DeepLabV3+ [101] |
| Context [192] | Generic | 59 | 10103 | – | 9637 | Variable | 2014 | 51.6% MIoU | CCL [115] |
| SBD [210] | Outdoor | 21 | 8498 | – | 2857 | Variable | 2011 | 82.1% MIoU | DeepLabv2+RWN [164] |
| RGB-D Object v2 [203] | Household / Warehouse objects | 51 | | 41877 | | $640 \times 480$ | 2014 | | |
| ScanNetv2 [199] | Indoor / 3D | 20 | | +1500 Scans | | Variable | 2018 | | |
| SegTrack v2 [191] | Generic / videos | 14 | | 976 Frames | | Variable | 2013 | 80.12% MIoU | RFCNet [149] |
| Sift-Flow [211] | Outdoor | 33 | 2488 | – | 200 | $256 \times 256$ | 2011 | 44.9% MIoU | Context-cNN [169] |
| Stanford | | | | | | | | | |
| Background [209] | Outdoor | 8 | | 715 | | $320 \times 240$ | 2009 | 65.7% MIoU | MCRNN [66] |
| 2D-3D [200] | Indoor / 2D-3D | 13 | | 70469 / 360° Scans | | $1080 \times 1080$ | 2017 | 49.9% fwIoU | Depth-CNN [215] |
| SUN Dataset | | | | | | | | | |
| 3D [201] | Indoor / 3D / video | – | | 19640 Frames | | $640 \times 80$ | 2013 | 58.5% IoU | LSTM-CF [216] |
| RGB-D [131] | Indoor / 2D-3D | 37 | 2666 | 2619 | 5050 | Variable | 2015 | 48.1% MIoU | CCL [115] |
| SYNTHIA [186] | Synthetic / Street view | 11 | | 13407 | | $960 \times 720$ | 2016 | 81.2% MIoU | RFCNet [149] |
| Youtube Dataset [204] | Objects / Video | 10 | | +10000 Frames / 126 Videos | | $480 \times 360$ | 2014 | 68.5% MIoU | Clockwork-FCN [144] |

semantic segmentation. Shotton et al. [13,198,203] datasets are not accessed for semantics, but they can be used for semantic segmentation. Huang et al. [184,199] datasets are not evaluated at all. All these datasets provide appropriate pixel-wise or point-wise labels.

### 3.2. Evaluation

We describe commonly used evaluation metrics for semantic segmentation. The overall performance of the semantic segmentation systems can be assessed in terms of accuracy, time, memory, and power consumption.

*Accuracy:* The accuracy of the semantic segmentation system is measure of the correctness of the segmentation or is the ratio of the correctly segmented area over the ground truth.

*Pixel wise Accuracy*: The ratio between the amount of correctly classified pixels and the total number of them. Confusion matrix terminology is used to describe the performance of a classification model.

Let $N_{cls}$ be the number of classes, $N_{xy}$ is the number of pixels which belong to class x and were labeled as class y. The confusion matrix reports the number of false positives ($N_{xy}$), false negatives ($N_{yx}$), true positives ($N_{xx}$), and true negatives ($N_{yy}$).

$$PixelAccuracy = \frac{\Sigma_{x=1}^{N_{cls}} N_{xx}}{\Sigma_{x=1}^{N_{cls}} \Sigma_{y=1}^{N_{cls}} N_{xy}} \tag{1}$$

The pixel-wise classification accuracy is not reliable for the real performance of a classifier, because it will yield misleading results if the data set is unbalanced (i.e., large regions which have one class or labeled images could have a more coarse labeling).

*Mean Accuracy:* The ratio of correct pixels is calculated in per-class basis and then averaged over the total number of classes $N_{cls}$.

$$MeanAccuracy = \frac{1}{N_{cls}} \Sigma_{x=1}^{N_{cls}} \frac{N_{xx}}{\Sigma_{y=1}^{N_{cls}} N_{xy}} \tag{2}$$

*Mean Intersection over Union (MIoU):* The ratio between the numbers of true positives $N_{xx}$, (intersection) over the sum of true positives $N_{xx}$, false negatives $N_{yx}$, false positives $N_{xy}$ (union). Intersection over Union is computed on a per-class basis and then averaged.

$$MIoU = \frac{1}{N_{cls}} \Sigma_{x=1}^{N_{cls}} \frac{N_{xx}}{\Sigma_{y=1}^{N_{cls}} N_{xy} + \Sigma_{y=1}^{N_{cls}} N_{yx} - N_{xx}} \tag{3}$$

The most widely used accuracy measuring strategy is MIoU, due to its easiness and simplicity.

*Frequency Weighted Intersection over Union (FWIoU)*

$$FWIoU = \frac{1}{\Sigma_{x=1}^{N_{cls}} \Sigma_{y=1}^{N_{cls}} N_{yx}} \Sigma_{x=1}^{N_{cls}} \frac{\Sigma_{y=1}^{N_{cls}} N_{xy} N_{xx}}{\Sigma_{y=1}^{N_{cls}} N_{xy} + \Sigma_{y=1}^{N_{cls}} N_{yx} - N_{xx}} \tag{4}$$

*Precision:* The relation between true positives $N_{xx}$, and all elements classified as positives

$$Precision = \frac{N_{xx}}{N_{xx} + N_{xy}} \tag{5}$$

*Recall:* measures how good all the positives are found.

$$Recall = \frac{N_{xx}}{N_{xx} + N_{yx}} \tag{6}$$

*Average Precision:* Mean precision at a set of eleven equal space recall levels (0.0, 0.1, 0.2..., 1)

*Mean Average Precision:* Mean of all the Average Precision values across all classes.

*Time, memory and power:* The memory and processing time of the system is highly dependent on hardware and the back-end implementation. The usage of hardware accelerators GPUs makes the processing time of these system very fast, however it consumes much of the memory and power. Most of the methods do not provide information, regarding time, memory and hardware, which is very crucial as these network models may be applied in (mobile systems, robotics, autonomous driving etc) where with limited power and memory, extremely accurate image segmentation would be required. Furthermore, these information can help researchers to estimate, make comparisons or choose methods depending on the application and requirement.

## 4. Analysis & discussion

We analyze some of the network models on the bases of their performance on datasets and their design structure to find out the reasons for their accomplishments. It is difficult to compare these methods due to the majority of them has been evaluated on very few datasets. Some methods used different metrics and also lack information about experimental setup (hardware, time, memory).

### 4.1. AdapNet [191]

- Achieves top score of 88.25% IoU on Freiburg Forest and 72.91% IoU on Synthia dataset. The network achieves the score of 69.39% IoU on cityscapes dataset.

The improvement can be credited to the highly representational multi-scale features learned by the model, which enable the segmentation of very distant objects present in Synthia and Cityscapes. AdapNet model approach is based on a mixture of convolutional neural network (CNN) experts (Convoluted Mixture of Deep Experts - CMoDE) and incorporates multiple modalities including appearance, depth and motion.

### 4.2. PSPNet [112]

- Achieves the best results on ADE20K with 44.8% IoU, promising results are obtained on cityscapes and Pascal VOC with 80.2% IoU and 85.4% IoU respectively.

PSPNet developed an effective optimization strategy for deep ResNet-101 [34] based on deeply supervised loss; two loss functions: main softmax loss to train the final classifier and auxiliary loss applied after the fourth stage, this helps optimizing the learning process. PSPNet applies multi scale testing, experiments different depths of pre-trained ResNet and data augmentation is performed.

### 4.3. FCCN [83]

- Achieves a top scores of 69.94% IoU on CamVid and score of 44.23% IoU on ADE20K dataset.

FCCN proposed a cost function that significantly improves the segmentation performance, very few researchers tried to modify cost function when training their models. FCCN calculates cost function on each pre output layer including the final output layer.

### 4.4. DeepLab V3 [100]

- Achieves score of 81.3% IoU on cityscapes.

Improvement mainly comes from changing hyper perimeter: Fine tuning batch normalization, varying batch size, larger crop size, changing output stride, multi scale inputs during inference, adding left-right flipped inputs, trained on 3475 finely and extra 20000 coarsely annotated images of cityscapes dataset. Furthermore, the use of ResNet-101 model which is pre-trained on ImageNet and JFT dataset, results in the second best score of 86.90 IoU on Pascal VOC.

### 4.5. DeepLab V3+ [101]

- Achieves 89.0% IoU on Pascal VOC and 82.1% IoU on cityscapes.

DeepLab V3+ is a modified version of DeepLab V3, adapted to output stride = 16 or 8 instead of 32. It is also adapted to Xception module, which further increased the performance.

### 4.6. DSSPN [43]

- Achieves top scores on COCO Stuff 38.9% IoU, 43.6% IoU on ADE20K, 58.6% IoU on Pascal Context and 45.01% IoU on Mapillary dataset.

DSSPN constructs a semantic neuron graph in which each neuron segments regions of one parent concept in a semantic concept hierarchy (combining labels from four datasets) and aims at recognizing between its child concepts. Instead of using a completely large semantic neural graph, DSSPN only activates relative small neural graph for each image during training, which makes DSSPN memory and computation efficient.

### 4.7. RFCNet [149]

- Achieves top scores of 81.20% IoU on SYNTHIA, 80.12% IoU on SegTrack and 69.84% IoU on DAVIS dataset.

The model uses different FCN architectures as a recurrent node to utilize temporal information, deconvolution layer for upsampling and supports skip architecture for finer segmentation. The use of temporal data is the reason for the boost of performance not just simply adding extra convolutional filters.

### 4.8. Adelaide Context CNN-CRF [169]

- Achieves score of 40.6% IoU on NYUDv2, 42.30% IoU on SUN-RGB, 78.00% IoU on Pascal VOC, 66.40% IoU on CIFAR-100, 71.60% IoU on Cityscapes, and 43.30% IoU on Pascal Context dataset.

The model uses CNN-based pairwise potential functions to capture semantic correlations between neighboring patches which improve the coarse-level prediction. The model applies FCN with sliding pyramid pooling, CNN contextual pairwise, boundary refinement (dense CRF method), and trained model with extra images from the COCO dataset to improve the overall performance of the model.

### 4.9. Clockwork-FCN [144]

- Achieves 68.50% IoU on Youtube Object, 68.40% IoU on Cityscapes, 28.90% IoU on NYUDv2 dataset.

The Clockwork-FCN uses different clock schedules; Fixed-rate clock reduces computation by assigning different rates to each stage such that later stages execute less often. Adaptive clockwork updates when the output score maps is predicted to change, thus reducing computation while maintaining accuracy.

### 4.10. Residual framework ResNet-38 [38]

- Achieves the highest score of 48.1% IoU on Pascal Context, 80.6% IoU on cityscapes and 43.43% IoU on ADE20K.

The model introduces residual units into ResNet (17 residual units for 101 layers ResNet) expanding it into a sufficiently large number of sub-networks. Each connection in residual unit shares same kernel sizes and numbers of channels, this results in improving model accuracy. ResNet-38 does not apply any multi-scale testing, model averaging or CRF based post-processing, except for the test set of ADE20K.

### 4.11. ESPNet [47]

- Efficient real-time segmentation network, achieves 60.2% IoU on cityscape, 40.0% IoU on Mapillary dataset with 0.364M parameters, 63.01% IoU on Pascal VOC test set with 0.364M parameters.

Efficient Spatial Pyramid (ESP) network is an efficient neural network in terms of speed and memory. ESP, based on factorized form of convolutions (point-wise convolution and spatial pyramid of dilated convolutions), reduces the number of parameters, memory, with large receptive field.

### 4.12. FCN-8s [77]

- Achieves the score of 77.46% IoU on Freiburg Forest, 67.20% IoU on PASCAL VOC, 65.30% IoU on CIFAR-10, 65.30% IoU on Cityscapes, 56.10% IoU on KITTI, 29.39% IoU on ADE20K, 35.10% IoU on PASCAL CONTEXT, 65.24% IoU on SYNTHIA, and 57.00% IoU on CamVid dataset.

The performance is increased by transferring pre-trained classifier weights, fusing different layer representations, and learning end-to-end on whole images.

### 4.13. DAG-RNN [72]

- Achieves 44.8% IoU on Sift-flow, 31.2% IoU on COCO stuff (171 classes) and 43.7% IoU on PASCAL Context dataset.

Segmentation network uses a pre-trained CNN with DAG-RNN, fusing low-level features with DAG-RNN. A new class weighted loss function proposed to control the classwise loss during training. The performance of segmentation network increases with increase in DAGs with DAG-RNN. Fully connected CRF is used, which further improves the performance of the network.

### 4.14. RefineNet [88]

- Achieves a score of 45.90% IoU on SUN-RGB, 46.50% IoU on NYUDv2 and 47.30% IoU on Pascal Context datasets. The results on Pascal VOC, cityscapes, and ADE20K datasets are 83.40% IoU, 73.60% IoU, and 40.70 % IoU respectively.

RefineNet applies data augmentation during training (random scaling, cropping and horizontal flipping of image), and multi-scale evaluation (average the predictions on the same image across different scales for the final prediction). Dense CRF method is used only for Pascal VOC.

### 4.15. Dilation10 [98]

- Achieves 67.60% IoU on PASCAL VOC, 67.10% IoU on Cityscapes, 32.31% IoU on ADE20K and 65.29% IoU on CamVid dataset.

The model is an adapted version of [69], replacing the pooling and convolutional layers of conv4/conv5 with two dilated convolution layers with dilation factors of 2 and 4 respectively. This leads to a decrease in the size of the network and its running time for real-time applications.

### 4.16. ResNet DUC+HDC [103]

- Achieves a score of 80.10% IoU on Cityscapes, 83.10% IoU on PASCAL VOC, 39.40% IoU on ADE20K dataset.

DUC provides the dense pixel-wise predictions, HDC uses arbitrary dilation rates which enlarge the receptive fields of the network. ResNet with different depths are experimented, data augmentation is performed (for cityscapes, each image of the training set is partitioned into twelve $800 \times 800$ patches making 35700 images). The model is trained using the combination of MS-COCO dataset, augmented PASCAL VOC 2012 training and trainval sets. ResNet DUC+HDC is also evaluated on KITTI dataset achieving the average precision of 92.88% for road segmentation using ResNet 101-DUC model, pre-trained from ImageNet during training.

### *4.17. ST-Dilation [145]*

- Achieves the score of 65.90% IoU on CamVid dataset. Model ST-FCN32s scores 50.60% IoU on Camvid dataset and Model ST-FCN8s scores 30.90% IoU on NYUDv2 dataset.

In STFCN model, no post processing required, the spatial temporal module is embedded on top of the final convolutional layer. LSTM blocks are used for inferring the relations between spatial features that provide valuable information and improve the accuracy of the segmentation. Furthermore, applying dilated convolutions for multi-scale contextual information archives better results.

### *4.18. STGRU (GRFP + Dilation) [151]*

- Achieves the score of 66.10 IoU on CamVid dataset. Model GRFP + Dilation scores 67.80% IoU and model GRFP + LRR-4x achieves the score of 72.80% IoU on Cityscapes dataset.

The model combines the power of both convolutional-gated architecture and spatial transformers (CNN). The model GRFP is trained with Dilation 10 [88] and LRR [70] network that improve performance for video. The model improves semantic video segmentation and labeling accuracy by propagating information from labeled video frames to nearby unlabeled frames with slight computation.

It can be noticed, that those methods which achieved the high performance results, are doing so due to the availability of large amount of labeled data. Extra training data is beneficial for increasing the accuracy of the model; several models used large datasets (merging two or three datasets) during training.

## 5. Open problems and challenges

In this section, we discuss some of the open problems and their possible solutions.

### *5.1. Open problems and possible solutions*

Techniques for semantic segmentation using deep neural networks (DNNs) are rapidly growing and the following problems are still needed to be addressed.

#### *5.1.1. Reducing complexity & computation*

The deep neural networks are not much suitable to be deployed on mobile platforms (e.g. embedded devices) that have limited resources because, DNN are highly memory demanding, time and power consuming. There is also problem with computational complexity that arises due to a great number of operations needed for inference. It is important to investigate how to reduce the complexity of the model to achieve high efficiency without loss of accuracy. Some CNN compression approaches have been proposed to deal with reducing complexity and computational cost. Wang et al. [217] proposed a method to excavate and decrease the redundancy in feature maps extracted from large number of filters in each layer of network. Kim et al. [218] proposed a one-shot whole network compression approach, that consists of three steps: Rank selection, Low-rank tensor decomposition, and fine tuning. Holliday et al. [219] applied model compression techniques to the problem of semantic segmentation. Caffe2 is a portable deep learning framework by Facebook, capable of training large models and allows to build machine learning applications for mobile systems. Compressing and accelerating DNN achieved lots of progress. However, there are some potential issues like; compression may cause loss in accuracy; decomposition operation; transfer information to convolutional filters is not suitable on some networks.

#### *5.1.2. Apply to adverse conditions*

There have been a few of network models which are applied in real challenging environmental conditions or to handle adverse conditions such as direct lighting, reflections from specular surfaces, varying seasons, fog or rain. Although, some CNN models used synthetic data together with real data to boost the performance of state-of-the-art methods for semantic segmentation on the challenging environmental conditions. However, using huge amounts of high-quality data from the real world so far remains indispensable. One possible solution is to use synthetic data with the real data. Apparently there are significant visual differences between the two data domains and to narrow this gap, domain adaptation technique may be used. Hoffman et al. [220] proposed an unsupervised domain adaptation method for transferring semantic segmentation FCNs across image domains. Zhang et al. [221] proposed a curriculum-style learning approach to minimize the domain gap. The authors in [222] proposed a domain Shift approach based on Generative Adversarial Network (GAN), which transfers the information of the target distribution to the learned embedding using a generator-discriminator pair.

#### *5.1.3. Need large and high quality labeled data*

The classification performance of DNNs and dataset size are positively correlated. Current state-of-the-art methods require high quality labeled data, which is not available on large-scale as they are time consuming and labour exhaustive. The effective solution to this problem would be to build large and high quality datasets, which seems hard to achieve. Therefore, the researchers rely on semi and weakly supervised methods making DNNs less reliant on the labeling of large datasets. These methods has considerably improved the semantic segmentation performance by using additional weak annotations either alone or in combination with a small number of strong annotations. However, they are far from fully supervised learning methods in terms of accuracy. Thus, this opens new challenges for improvement.

#### *5.1.4. Overfitting*

As mentioned before, DNNs are data hungry and they do not perform well unless they are fed with large datasets. Majority of the available datasets are relatively small, so as DNN models become very complex to capture all the useful information necessary to solve a problem. The model may run risk of “Overfitting” with limited amount of data. Overfitting occurs when the gap between the training error and test error is too large. Regularization techniques help in overcoming this problem. Regularization is any modification we make to a learning algorithm that is intended to reduce its generalization error but not its training error [60]. Several of these methods are applied in DNNs to prevent overfitting such as L1 and L2 regularization, Lp norm, dropout, dropConnect. Data Augmentation is also used to reduce overfitting (e.g. increasing the size of the training data - image rotating, flipping, scaling, shifting). However, the regularization may increase training time (e.g. using dropout increases the training time by 2x or 3x than a standard neural network of the same architecture) and there is no standard for regularizing CNNs. Introducing better or improved regularization method would be an interesting direction for future work.

*5.1.5. Segmentation in real-time*

Real-time semantic segmentation without loosing to much accuracy is of great importance, as it can be useful in autonomous driving, robot interaction, mobile computing where running time is critical to evaluate the performance of the system. DNN methods for semantic segmentation are more focused on accuracy then speed. Majority of the methods are far away from real-time segmentation. One possible solution to the problem could be performing convolutions in an efficient manner. Several works have aimed at developing efficient architectures that can run in real-time based on convolution factorization (disintegrate convolutional operation into multiple steps). Some computationally efficient modules for convolution are introduced. For example, Inception [27], Xception [31], ResNet [34], ASP [99], ESP [47]; ShuffleNet [223] and MobileNet [224], are using grouped and depth-wise convolutions. Another possible solution could be to apply network compression using different techniques (e.g. parameter pruning and sharing [225], low-rank factorization and sparsity [226] etc.) to reduce the size of the network. However, real-time semantic segmentation still lacks higher accuracy and new methods and approaches must be developed to work-out between runtime and accuracy.

*5.1.6. Video / 3D segmentation*

DNNs have been successfully applied for semantic segmentation of 2D images while not much for 3D images and on videos despite their significance. Several video and 3D network models for semantic segmentation have been proposed over the years and progress has been made but some challenges still exist. The lack of large datasets of 3D images and sequence images (videos) make it difficult to progress on 3D and video semantic segmentation. 3D networks are computationally expensive when dealing with high resolution and complex scenes (large number of classes). In 3D semantic segmentation task, using 3D Point cloud information is very effective. Zhang et al. [227] proposed an efficient large-scale point cloud segmentation method, in which 2D images with 3D point clouds are fused into CNN to segment complex 3D urban scenes. The authors in [228,229] proposed methods for direct semantic labeling of 3D pointclouds with spectral information. However, 3D segmentation methods face many challenges as compared to 2D segmentation, i.e., High complexity, computational cost, slow processing and most important lack of 3D datasets. In video semantic segmentation, two approaches can be useful, one to improve computational cost (by reducing latency); The authors in [144,230] proposed designed schedule frameworks which reduce the overall cost and maximum latency of video semantic segmentation. However, these approaches are far away to meet the latency requirements in real-time applications. The second approach is to improve accuracy (by exploiting temporal continuity - temporal features and temporal correlations between video frames). Several methods [145,146,148] have been proposed using temporal information with spatial information for increasing the accuracy of pixel labeling.

## 6. Conclusions

In this paper, we have provided a comprehensive survey of deep learning techniques used for semantic segmentation.

The surveyed methods have been categorized in ten classes, according to the common concept underlaying their architectures. We have also provided a summary of these methods stating, for each of them, the main idea, the origin of its architecture, testing benchmarks, code availability and the year of publication.

Thirty five datasets on which these methods have been applied, have been reported and described in details showing their environment nature, number of classes, resolution, number of the images and the method which achieved the best performance on each till date to the best of our knowledge.

We have mainly analyzed the design and performance of some of these methods which reported that had achieved high scores. The goal was to find out how they do so.

We have also discussed some of the open problems and tried to suggest some of possible solutions.

This survey had shown that there is much scope of improvement in terms of accuracy, speed and complexity. So, our future work, would be to take some of these methods and develop a new one by enhancement of the weaknesses and/or combination of the merits.

## Acknowledgments

The authors express their gratitude to University Technology Belfort-Montbeliard and Higher Education Commission of Pakistan for providing the support and necessary requirement for completion of work. The authors would also like to acknowledge Zhi Yan and Abdellatif El Idrissi for helpful discussions.

## Appendix A

Table 13.

**Table 13**
Links to the source codes.

| Network model | Code link | Network model | Code link |
|---|---|---|---|
| Inception [27] | https://github.com/Microsoft/CNTK/tree/master/Examples/Image/Classification/GoogLeNet | RSIS [75] | https://github.com/imatge-upc/rsis |
| BN-Inception [28] | | FC-DenseNet [51] | https://github.com/SimJeg/FC-DenseNet |
| Inception V2, V3 [29] | | ConvDeconvNet [79] | https://github.com/HyeonwooNoh/DeconvNet |
| Inception V4 [30] | https://github.com/titu1994/Inception-v4 | SegNet [80] | https://github.com/alexgkendall/caffe-segnet |
| Xception [31] | https://github.com/kwotsin/TensorFlow-Xception | FCN [77] | https://github.com/shelhamer/fcn.berkeleyvision.org |
| VGGNet [33] | https://github.com/machrisaa/tensorflow-vgg | SMSNet [84] | https://github.com/JohanVer/SMSnet |
| ResNet [34] | https://github.com/KaimingHe/deep-residual-networks | ICNet [87] | https://github.com/hszhao/ICNet |
| ResNet-38 [38] | https://github.com/itijyou/ademxapp | RefineNet [88] | https://github.com/guosheng/refinenet |
| ResNeXt [41] | https://github.com/facebookresearch/ResNeXt | RDFNET [89] | https://github.com/SeongjinPark/RDFNet/blob/master |
| INPLACE-ABN [42] | https://github.com/mapillary/inplace_abn | G-FRNet [90] | https://github.com/mrochan/gfrnet |
| FRRN [44] | https://github.com/TobyPDE/FRRN | LRN [91] | https://github.com/golnazghiasi/LRR |
| ENet [45] | https://github.com/TimoSaemann/ENet | DeepLab [97] | https://bitbucket.org/deeplab/deeplab-public |
| ERFNet [46] | https://github.com/Eromera/erfnet | DeepLabV2 [99] | https://bitbucket.org/aquariusjay/deeplab-public-ver2 |
| ESPNet [47] | https://github.com/sacmehta/ESPNet | Dilation [98] | https://github.com/tensorflow/models/tree/master/research/deeplab |
| R-CNN [52] | https://github.com/rbgirshick/rcnn | DeepLabV3+ [101] | https://github.com/fyu/dilation |
| Fast R-CNN [54] | https://github.com/rbgirshick/fast-rcnn | HDC [103] | https://github.com/TuSimple/TuSimple-DUC |
| Faster R-CNN [55] | https://github.com/ShaoqingRen/faster_rcnn | DRN [104] | https://github.com/fyu/drn |
| Mask R-CNN [56] | https://github.com/matterport/Mask_RCNN | PSPNet [112] | https://github.com/hszhao/PSPNet |
| FPN [57] | https://github.com/unsky/FPN | DenseCRF [160] | https://github.com/lucasb-eyer/pydensecrf |
| DecoupledNet [122] | https://github.com/HyeonwooNoh/DecoupledNet | GCF [163] | https://github.com/siddharthachandra/gcrf |
| WSSL [123] | https://bitbucket.org/deeplab/deeplab-public | ConvCRF [164] | https://github.com/MarvinTeichmann/ConvCRF |
| Semi-Adv [127] | https://github.com/hfslyc/AdvSemiSeg | BNN [171] | https://github.com/MPI-IS/bilateralNN |
| DSRG [132] | https://github.com/speedinghzl/DSRG | Clockwork [144] | https://github.com/shelhamer/clockwork-fcn |
| MCG GrabCut+ [137] | https://github.com/philferriere/tfwss | STFCN [145] | https://github.com/MohsenFayyaz89/STFCN |
| ReSeg [64] | https://github.com/fvisin/reseg | FSO [147] | https://bitbucket.org/infinitei/videoparsing |

## References

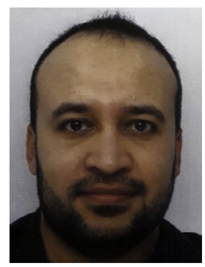

[1] D.D. Cox, T. Dean, Neural networks and neuroscience-inspired computer vision, Curr. Biol. 24 (18) (2014) R921–R929.
[2] B. Li, S. Liu, W. Xu, W. Qiu, Real-time object detection and semantic segmentation for autonomous driving, in: Proceedings of the MIPPR 2017 (Automatic Target Recognition and Navigation), 10608, International Society for Optics and Photonics, 2018, p. 106080P.
[3] Y.-H. Tseng, S.-S. Jan, Combination of computer vision detection and segmentation for autonomous driving, in: Proceedings of the IEEE/ION Position, Location and Navigation Symposium (PLANS), IEEE, 2018, pp. 1047–1052.
[4] Y. Zhang, H. Chen, Y. He, M. Ye, X. Cai, D. Zhang, Road segmentation for all-day outdoor robot navigation, Neurocomputing 314 (2018) 316–325.
[5] X. Tao, D. Zhang, W. Ma, X. Liu, D. Xu, Automatic metallic surface defect detection and recognition with convolutional neural networks, Appl. Sci. 8 (9) (2018) 1575.
[6] R. Kemker, C. Salvaggio, C. Kanan, Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning, ISPRS J. Photogramm. Remote Sens. 145 (2018) 60–77.
[7] Y. Ji, H. Zhang, K.-K. Tseng, T.W. Chow, Q.J. Wu, Graph model-based salient object detection using objectness and multiple saliency cues, Neurocomputing 323 (2019) 188–202.
[8] Y. Ji, H. Zhang, Q.J. Wu, Salient object detection via multi-scale attention CNN, Neurocomputing 322 (2018) 130–140.
[9] A. Milioto, P. Lottes, C. Stachniss, Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in CNNS, in: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2018, pp. 2229–2235.
[10] H. Zhang, Y. Sun, L. Liu, X. Wang, L. Li, W. Liu, Clothingout: a category-supervised GAN model for clothing segmentation and retrieval, Neural Comput. Appl. (2018) 1–12, doi:10.1007/s00521-018-3691-y.
[11] F. Jiang, A. Grigorev, S. Rho, Z. Tian, Y. Fu, W. Jifara, K. Adil, S. Liu, Medical image semantic segmentation based on deep learning, Neural Comput. Appl. 29 (5) (2018) 1257–1265.
[12] J. Shotton, M. Johnson, R. Cipolla, Semantic texton forests for image categorization and segmentation, in: Proceedings of the Computer Vision and Pattern Recognition, IEEE, 2008, pp. 1–8.
[13] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, A. Blake, Real-time human pose recognition in parts from single depth images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2011, pp. 1297–1304.
[14] Semantic segmentation deep learning review, http://blog.qure.ai/notes/semantic-segmentation-deep-learning-review.
[15] H. Zhu, F. Meng, J. Cai, S. Lu, Beyond pixels: a comprehensive survey from bottom-up to semantic image segmentation and cosegmentation, J. Vis. Commun. Image Represent. 34 (2016) 12–27.
[16] M. Thoma, A survey of semantic segmentation, CoRR (2016). abs/1602.06541.
[17] J. Niemeijer, P.P. Fouopi, S. Knake-Langhorst, E. Barth, A review of neural network based semantic segmentation for scene understanding in context of the self driving car, Proceedings of the BioMedTec Studierendentagung, 2017.
[18] Y. Guo, Y. Liu, T. Georgiou, M.S. Lew, A review of semantic segmentation using deep neural networks, Int. J. Multimedia Inf. Retr. 7 (2) (2018) 87–93.
[19] Q. Geng, Z. Zhou, X. Cao, Survey of recent progress in semantic image segmentation with CNNS, Sci. China Inf. Sci. 61 (5) (2018) 051101.
[20] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, J. Garcia-Rodriguez, A review on deep learning techniques applied to semantic segmentation, Soft Comput. 70 (2017) 41–65.
[21] H. Yu, Z. Yang, L. Tan, Y. Wang, W. Sun, M. Sun, Y. Tang, Methods and datasets on semantic segmentation: a review, Neurocomputing 304 (2018) 82–103.
[22] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.
[23] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of the Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[24] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Proceedings of the European Conference on Computer Vision, Springer, 2014, pp. 818–833.
[25] M. Lin, Q. Chen, S. Yan, Network in network, CoRR (2013). abs/1312.4400.
[26] F. Rosenblatt, Principles of neurodynamics. perceptrons and the theory of brain mechanisms, Technical Report, Cornell Aeronautical Lab Inc Buffalo NY, 1961.
[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[28] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, ICML 2015, 2015.
[29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.
[30] C. Szegedy, S. Ioffe, V. Vanhoucke, A.A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning., in: Proceedings of the AAAI, 4, 2017, p. 12.
[31] F. Chollet, Xception: deep learning with depthwise separable convolutions, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1800–1807, doi:10.1109/CVPR.2017.195.
[32] F. Mamalet, C. Garcia, Simplifying convnets for fast learning, in: Proceedings of the International Conference on Artificial Neural Networks, Springer, 2012, pp. 58–65.
[33] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, CoRR (2015). abs/1409.1556.
[34] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[35] G.E. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 1527–1554.
[36] G.E. Hinton, Deep belief networks, Scholarpedia 4 (5) (2009) 5947.
[37] W. Byeon, T.M. Breuel, F. Raue, M. Liwicki, Scene labeling with lstm recurrent neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3547–3555.
[38] Z. Wu, C. Shen, A. van den Hengel, Wider or deeper: Revisiting the resnet model for visual recognition, CoRR (2016). abs/1611.10080.
[39] A. Casanova, G. Cucurull, M. Drozdzal, A. Romero, Y. Bengio, On the iterative refinement of densely connected representation levels for semantic segmentation (2018) arXiv preprint arXiv: 1804.11332.
[40] A. Valada, J. Vertens, A. Dhall, W. Burgard, Adapnet: Adaptive semantic segmentation in adverse environmental conditions, in: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2017, pp. 4644–4651.
[41] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 5987–5995.
[42] S.R. Bulò, L. Porzi, P. Kontschieder, In-place activated batchnorm for memory-optimized training of DNNs, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018) 5639–5649.
[43] X. Liang, H. Zhou, E. Xing, Dynamic-structured semantic propagation network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 752–761.
[44] T. Pohlen, A. Hermans, M. Mathias, B. Leibe, Full-resolution residual networks for semantic segmentation in street scenes, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3309–3318.
[45] A. Paszke, A. Chaurasia, S. Kim, E. Culurciello, Enet: A deep neural network architecture for real-time semantic segmentation, CoRR (2016). abs/1606.02147.
[46] E. Romera, J.M. Alvarez, L.M. Bergasa, R. Arroyo, Erfnet: Efficient residual factorized convnet for real-time semantic segmentation, IEEE Trans. Intell. Transp. Syst. 19 (1) (2018) 263–272.
[47] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, H. Hajishirzi, ESPNet: Efficient spatial pyramid of dilated convolutions for semantic segmentation, ECCV, 2018.
[48] L. Deng, M. Yang, H. Li, T. Li, B. Hu, C. Wang, Restricted deformable convolution based road scene semantic segmentation using surround view cameras, CoRR (2018). abs/1801.00708.
[49] G. Huang, Y. Sun, Z. Liu, D. Sedra, K.Q. Weinberger, Deep networks with stochastic depth, in: Proceedings of the European Conference on Computer Vision, Springer, 2016, pp. 646–661.
[50] A. Veit, M.J. Wilber, S. Belongie, Residual networks behave like ensembles of relatively shallow networks, in: Proceedings of the Advances in Neural Information Processing Systems, 2016, pp. 550–558.
[51] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, Y. Bengio, The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2017, pp. 1175–1183.
[52] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587.
[53] J.R. Uijlings, K.E. Van De Sande, T. Gevers, A.W. Smeulders, Selective search for object recognition, Int. J. Comput. Vis. 104 (2) (2013) 154–171.
[54] R. Girshick, Fast R-CNN, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.
[55] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, in: Proceedings of the Advances in Neural Information Processing Systems, 2015, pp. 91–99.
[56] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), IEEE, 2017, pp. 2980–2988.
[57] T.-Y. Lin, P. Dollár, R.B. Girshick, K. He, B. Hariharan, S.J. Belongie, Feature pyramid networks for object detection., in: Proceedings of the CVPR, 1, 2017, p. 4.
[58] S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, Path aggregation network for instance segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8759–8768.
[59] J. Hosang, R. Benenson, P. Dollár, B. Schiele, What makes for effective detection proposals? IEEE Trans. Pattern Anal. Mach. Intell. 38 (4) (2016) 814–830.
[60] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016. http://www.deeplearningbook.org
[61] A. Graves, A.-r. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2013, pp. 6645–6649.
[62] F. Gao, L. Wu, L. Zhao, T. Qin, X. Cheng, T.-Y. Liu, Efficient sequence learning with group recurrent networks, in: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (Long Papers), 1, 2018, pp. 799–808.
[63] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[64] F. Visin, M. Ciccone, A. Romero, K. Kastner, K. Cho, Y. Bengio, M. Matteucci, A. Courville, Reseg: A recurrent neural network-based model for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 41–48.
[65] F. Visin, K. Kastner, K. Cho, M. Matteucci, A.C. Courville, Y. Bengio, ReNet: A recurrent neural network based alternative to convolutional networks, CoRR (2015). abs/1505.00393.
[66] H. Fan, X. Mei, D. Prokhorov, H. Ling, Multi-level contextual RNNS with attention model for scene labeling, IEEE Trans. Intell. Transp. Syst. 99 (2018) 1–11.
[67] L.-C. Chen, Y. Yang, J. Wang, W. Xu, A.L. Yuille, Attention to scale: scale-aware semantic image segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3640–3649.
[68] P.H. Pinheiro, R. Collobert, Recurrent convolutional neural networks for scene labeling, in: Proceedings of the 31st International Conference on Machine Learning (ICML), 2014. EPFL-CONF-199822
[69] B. Shuai, Z. Zuo, B. Wang, G. Wang, Dag-recurrent neural networks for scene labeling, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3620–3629.
[70] H. Fan, H. Ling, Dense recurrent neural networks for scene labeling, CoRR (2018). abs/1801.06831.
[71] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks., in: Proceedings of the CVPR, 1, 2017, p. 3.
[72] B. Shuai, Z. Zuo, B. Wang, G. Wang, Scene segmentation with dag-recurrent neural networks, IEEE Trans. Pattern Anal. Mach. Intell. 40 (6) (2018) 1480–1493.
[73] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder–decoder for statistical machine translation, EMNLP 2014, 2014.
[74] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[75] A. Salvador, M. Bellver, M. Baradad, F. Marques, J. Torres, X. Giro-i-Nieto, Recurrent Neural Networks for Semantic Instance Segmentation (2017) arXiv: 1712.00617.
[76] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, W.-c. Woo, Convolutional LSTM network: a machine learning approach for precipitation nowcasting, in: Proceedings of the Advances in Neural Information Processing Systems, 2015, pp. 802–810.
[77] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[78] Y. Wang, J. Liu, Y. Li, J. Yan, H. Lu, Objectness-aware semantic segmentation, in: Proceedings of the ACM on Multimedia Conference, ACM, 2016, pp. 307–311.
[79] H. Noh, S. Hong, B. Han, Learning deconvolution network for semantic segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1520–1528.
[80] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: a deep convolutional encoder-decoder architecture for image segmentation, CoRR 39 (12) (2015) 2481–2495. abs/1511.00561.
[81] J. Fu, Y. Wang, H. Lu, Stacked deconvolutional network for semantic segmentation, CoRR (2017). abs/1708.04943.
[82] G. Nanfack, A. Elhassouny, R.O.H. Thami, Squeeze-segnet: a new fast deep convolutional neural network for semantic segmentation, in: Proceedings of the Tenth International Conference on Machine Vision (ICMV), 10696, International Society for Optics and Photonics, 2018, p. 106962O.
[83] T. Yang, Y. Wu, J. Zhao, L. Guan, Semantic segmentation via highly fused convolutional network with multiple soft cost functions, Cognit. Syst. Res. 53 (2018) 20–30.
[84] J. Vertens, A. Valada, W. Burgard, Smsnet: Semantic motion segmentation using deep convolutional neural networks, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2017, pp. 582–589.
[85] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, T. Brox, Flownet 2.0: Evolution of optical flow estimation with deep networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2, 2017, p. 6.
[86] P. Bilinski, V. Prisacariu, Dense decoder shortcut connections for single-pass semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6596–6605.
[87] H. Zhao, X. Qi, X. Shen, J. Shi, J. Jia, Icnet for real-time semantic segmentation on high-resolution images, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 405–420.
[88] G. Lin, A. Milan, C. Shen, I.D. Reid, Refinenet: Multi-path refinement networks for high-resolution semantic segmentation., in: Proceedings of the CVPR, 1, 2017, p. 5.
[89] S. Lee, S.-J. Park, K.-S. Hong, Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), IEEE, 2017, pp. 4990–4999.
[90] M.A. Islam, M. Rochan, N.D. Bruce, Y. Wang, Gated feedback refinement network for dense image labeling, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017a, pp. 4877–4885.
[91] M.A. Islam, S. Naha, M. Rochan, N.D.B. Bruce, Y. Wang, Label refinement network for coarse-to-fine semantic segmentation, CoRR (2017). abs/1703.00551.
[92] G. Ghiasi, C.C. Fowlkes, Laplacian pyramid reconstruction and refinement for semantic segmentation, in: Proceedings of the European Conference on Computer Vision, Springer, 2016, pp. 519–534.
[93] F.J. Huang, Y.-L. Boureau, Y. LeCun, et al., Unsupervised learning of invariant feature hierarchies with applications to object recognition, in: Proceedings of

the IEEE Conference on Computer Vision and Pattern Recognition CVPR, IEEE, 2007, pp. 1–8.
[94] P.J. Burt, E.H. Adelson, The Laplacian pyramid as a compact image code, in: Proceedings of the Readings in Computer Vision, Elsevier, 1987, pp. 671–679.
[95] Y. Wu, T. Yang, J. Zhao, L. Guan, J. Li, Fully combined convolutional network with soft cost function for traffic scene parsing, in: Proceedings of the International Conference on Intelligent Computing, Springer, 2017, pp. 725–731.
[96] F.N. Iandola, M.W. Moskewicz, K. Ashraf, S. Han, W.J. Dally, K. Keutzer, SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and $<0.5$MB model size, CoRR (2017). abs/1602.07360.
[97] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Semantic image segmentation with deep convolutional nets and fully connected CRFs, CoRR (2015). abs/1412.7062.
[98] F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions, CoRR (2015). abs/1511.07122.
[99] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFS, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2018) 834–848.
[100] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, CoRR (2017). abs/1706.05587.
[101] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, ECCV, 2018.
[102] M. Treml, J. Arjona-Medina, T. Unterthiner, R. Durgesh, F. Friedmann, P. Schuberth, A. Mayr, M. Heusel, M. Hofmarcher, M. Widrich, et al., Speeding up semantic segmentation for autonomous driving, in: Proceedings of the ML-ITS, NIPS Workshop, 2016.
[103] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, G.W. Cottrell, Understanding convolution for semantic segmentation, 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) (2018) 1451–1460.
[104] F. Yu, V. Koltun, T.A. Funkhouser, Dilated residual networks., in: Proceedings of the CVPR, 2, 2017, p. 3.
[105] Z. Wu, C. Shen, A. van den Hengel, High-performance semantic segmentation using very deep fully convolutional networks, CoRR (2016). abs/1604.04339.
[106] M. Holschneider, R. Kronland-Martinet, J. Morlet, P. Tchamitchian, A real–time algorithm for signal analysis with the help of the wavelet transform, in: Wavelets, Springer, 1990, pp. 286–297.
[107] J.M. Alvarez, Y. LeCun, T. Gevers, A.M. Lopez, Semantic road segmentation via multi-scale ensembles of learned features, in: Proceedings of the European Conference on Computer Vision, Springer, 2012, pp. 586–595.
[108] C. Farabet, C. Couprie, L. Najman, Y. LeCun, Learning hierarchical features for scene labeling, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1915–1929.
[109] C. Couprie, C. Farabet, L. Najman, Y. Lecun, Indoor semantic segmentation using depth information, in: First International Conference on Learning Representations (ICLR 2013), Scottsdale, AZ, United States, 2013, pp. 1–8. https://hal.archives-ouvertes.fr/hal-00805105.
[110] S. Liu, X. Qi, J. Shi, H. Zhang, J. Jia, Multi-scale patch aggregation (mpa) for simultaneous detection and segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3141–3149.
[111] B. Hariharan, P. Arbeláez, R. Girshick, J. Malik, Hypercolumns for object segmentation and fine-grained localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 447–456.
[112] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2881–2890.
[113] D.M. Vo, S.-W. Lee, Semantic image segmentation using fully convolutional neural networks with multi-scale images and multi-scale dilated convolutions, Multimedia Tools Appl. 77 (14) (2018) 18689–18707.
[114] I. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, Y. Bengio, Maxout networks, in: S. Dasgupta, D. McAllester (Eds.), Proceedings of the 30th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol. 28, PMLR, Atlanta, Georgia, USA, 2013, pp. 1319–1327. http://proceedings.mlr.press/v28/goodfellow13.html.
[115] H. Ding, X. Jiang, B. Shuai, A.Q. Liu, G. Wang, Context contrasted feature and gated multi-scale aggregation for scene segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2393–2402.
[116] D. Lin, G. Chen, D. Cohen-Or, P.-A. Heng, H. Huang, Cascaded feature network for semantic segmentation of rgb-d images, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), IEEE, 2017, pp. 1320–1328.
[117] M. Mostajabi, P. Yadollahpour, G. Shakhnarovich, Feedforward semantic segmentation with zoom-out features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3376–3385.
[118] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Süsstrunk, et al., Slic superpixels compared to state-of-the-art superpixel methods, IEEE Trans. Pattern Anal. Mach. Intell. 34 (11) (2012) 2274–2282.
[119] D. Pathak, E. Shelhamer, J. Long, T. Darrell, Fully convolutional multi-class multiple instance learning., CoRR (2014). abs/1412.7144.
[120] P.O. Pinheiro, R. Collobert, From image-level to pixel-level labeling with convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1713–1721.
[121] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge university press, 2004.
[122] S. Hong, H. Noh, B. Han, Decoupled deep neural network for semi-supervised semantic segmentation, in: Proceedings of the Advances in Neural Information Processing Systems, 2015, pp. 1495–1503.
[123] G. Papandreou, L.-C. Chen, K.P. Murphy, A.L. Yuille, Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1742–1750.
[124] Z. Huang, X. Wang, J. Wang, W. Liu, J. Wang, Weakly-supervised semantic segmentation network with deep seeded region growing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7014–7023.
[125] P. Luo, G. Wang, L. Lin, X. Wang, Deep dual learning for semantic image segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017, pp. 21–26.
[126] N. Souly, C. Spampinato, M. Shah, Semi and weakly supervised semantic segmentation using generative adversarial network, CoRR (2017). abs/1703.09695.
[127] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, M.-H. Yang, Adversarial learning for semi-supervised semantic segmentation, BMVC, 2018.
[128] D. Barnes, W. Maddern, I. Posner, Find your own way: Weakly-supervised segmentation of path proposals for urban autonomy, in: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2017, pp. 203–210.
[129] F. Saleh, M.S. Aliakbarian, M. Salzmann, L. Petersson, S. Gould, J.M. Alvarez, Built-in foreground/background prior for weakly-supervised semantic segmentation, in: Proceedings of the European Conference on Computer Vision, Springer, 2016, pp. 413–432.
[130] F.S. Saleh, M.S. Aliakbarian, M. Salzmann, L. Petersson, J.M. Alvarez, S. Gould, Incorporating network built-in priors in weakly-supervised semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 40 (6) (2018) 1382–1396.
[131] S. Saito, T. Kerola, S. Tsutsui, Superpixel clustering with deep features for unsupervised road segmentation, CoRR (2017). abs/1711.05998.
[132] Z. Huang, X. Wang, J. Wang, W. Liu, J. Wang, Weakly-supervised semantic segmentation network with deep seeded region growing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7014–7023.
[133] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, T.S. Huang, Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7268–7277.
[134] L. Ye, Z. Liu, Y. Wang, Learning semantic segmentation with diverse supervision, 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) (2018) 1461–1469.
[135] J. Dai, K. He, J. Sun, Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1635–1643.
[136] Z. Wu, C. Shen, A. van den Hengel, Bridging category-level and instance-level semantic image segmentation, CoRR (2016). abs/1605.06885.
[137] A. Khoreva, R. Benenson, J.H. Hosang, M. Hein, B. Schiele, Simple does it: Weakly supervised instance and semantic segmentation., in: Proceedings of the CVPR, 1, 2017, p. 3.
[138] O. Maron, T. Lozano-Pérez, A framework for multiple-instance learning, in: Proceedings of the Advances in Neural Information Processing Systems, 1998, pp. 570–576.
[139] A. Khoreva, R. Benenson, J.H. Hosang, M. Hein, B. Schiele, Simple does it: Weakly supervised instance and semantic segmentation., in: Proceedings of the CVPR, 1, 2017, p. 3.
[140] C. Rother, V. Kolmogorov, A. Blake, Grabcut: Interactive foreground extraction using iterated graph cuts, in: Proceedings of the ACM Transactions on Graphics (TOG), 23, ACM, 2004, pp. 309–314.
[141] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929.
[142] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Proceedings of the Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.
[143] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, J. Feng, Y. Zhao, S. Yan, STC: a simple to complex framework for weakly-supervised semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (11) (2017) 2314–2320.
[144] E. Shelhamer, K. Rakelly, J. Hoffman, T. Darrell, Clockwork convnets for video semantic segmentation, in: Proceedings of the European Conference on Computer Vision, Springer, 2016, pp. 852–868.
[145] M. Fayyaz, M.H. Saffar, M. Sabokrou, M. Fathy, R. Klette, STFCN: Spatio-temporal FCN for semantic video segmentation, CoRR (2016). abs/1608.05971.
[146] Y. He, W.-C. Chiu, M. Keuper, M. Fritz, S. Campus, STD2P: RGBD semantic segmentation using spatio-temporal data-driven pooling., in: Proceedings of the CVPR, 2017, pp. 7158–7167.
[147] A. Kundu, V. Vineet, V. Koltun, Feature space optimization for semantic video segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3168–3175.
[148] Z. Qiu, T. Yao, T. Mei, Learning deep spatio-temporal dependence for semantic video segmentation, IEEE Trans. Multimedia 20 (4) (2018) 939–949.
[149] M. Siam, S. Valipour, M. Jagersand, N. Ray, Convolutional gated recurrent networks for video segmentation, in: Proceedings of the IEEE International Conference on Image Processing (ICIP), IEEE, 2017, pp. 3090–3094.

[150] F. Saleh, M.S.A. Akbarian, M. Salzmann, L. Petersson, J.M. Alvarez, Bringing background into the foreground: making all classes equal in weakly-supervised video semantic segmentation., in: Proceedings of the ICCV, 2017, pp. 2125–2135.
[151] D. Nilsson, C. Sminchisescu, Semantic video segmentation by gated recurrent flow propagation, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018) 6819–6828.
[152] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4489–4497.
[153] M.S. Pavel, H. Schulz, S. Behnke, Object class segmentation of RGB-D video using recurrent convolutional neural networks, Neural Netw. 88 (2017) 105–113.
[154] E.E. Yurdakul, Y. Yemez, Semantic segmentation of RGBD videos with recurrent fully convolutional neural networks., in: Proceedings of the ICCV Workshops, 2017, pp. 367–374.
[155] N. Ballas, L. Yao, C.J. Pal, A.C. Courville, Delving deeper into convolutional networks for learning video representations, CoRR (2015). abs/1511.06432.
[156] J. Koutnik, K. Greff, F. Gomez, J. Schmidhuber, A clockwork RNN, in: E.P. Xing, T. Jebara (Eds.), Proceedings of the 31st International Conference on Machine Learning, Proceedings of Machine Learning Research, vol. 32, PMLR, Bejing, China, 2014, pp. 1863–1871. http://proceedings.mlr.press/v32/koutnik14.html.
[157] S. Chandra, C. Camille, I. Kokkinos, Deep spatio-temporal random fields for efficient video segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8915–8924.
[158] S. Chandra, I. Kokkinos, Fast, exact and multi-scale inference for semantic image segmentation with deep gaussian CRFS, in: Proceedings of the European Conference on Computer Vision, Springer, 2016, pp. 402–418.
[159] A. Arnab, S. Zheng, S. Jayasumana, B. Romera-Paredes, M. Larsson, A. Kirillov, B. Savchynskyy, C. Rother, F. Kahl, P.H.S. Torr, Conditional random fields meet deep neural networks for semantic segmentation: combining probabilistic graphical models with deep learning for structured prediction, IEEE Signal Process. Mag. 35 (1) (2018) 37–52.
[160] P. Krähenbühl, V. Koltun, Efficient inference in fully connected CRFS with gaussian edge potentials, in: Proceedings of the Advances in Neural Information Processing Systems, 2011, pp. 109–117.
[161] A. Adams, J. Baek, M.A. Davis, Fast high-dimensional filtering using the permutohedral lattice, in: Proceedings of the Computer Graphics Forum, 29, Wiley Online Library, 2010, pp. 753–762.
[162] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, P.H. Torr, Conditional random fields as recurrent neural networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1529–1537.
[163] R. Vemulapalli, O. Tuzel, M.-Y. Liu, R. Chellapa, Gaussian conditional random field network for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3224–3233.
[164] M.T. Teichmann, R. Cipolla, Convolutional CRFS for semantic segmentation, arXiv:1805.04777 (2018).
[165] A. Arnab, S. Jayasumana, S. Zheng, P.H. Torr, Higher order conditional random fields in deep neural networks, in: Proceedings of the European Conference on Computer Vision, Springer, 2016, pp. 524–540.
[166] F. Shen, R. Gan, S. Yan, G. Zeng, Semantic segmentation via structured patch prediction, context CRF and guidance CRF, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 8, 2017.
[167] Z. Liu, X. Li, P. Luo, C.-C. Loy, X. Tang, Semantic image segmentation via deep parsing network, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1377–1385.
[168] G. Lin, C. Shen, I. Reid, A. van den Hengel, Deeply learning the messages in message passing inference, in: C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, R. Garnett (Eds.), Proceedings of the Advances in Neural Information Processing Systems 28, Currant Associates, Inc., 2015, pp. 361–369.
[169] G. Lin, C. Shen, A. Van Den Hengel, I. Reid, Efficient piecewise training of deep structured models for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3194–3203.
[170] J. Jiang, Z. Zhang, Y. Huang, L. Zheng, Incorporating depth into both cnn and crf for indoor semantic segmentation, in: Proceedings of the 8th IEEE International Conference on Software Engineering and Service Science (ICSESS), IEEE, 2017, pp. 525–530.
[171] V. Jampani, M. Kiefel, P.V. Gehler, Learning sparse high dimensional filters: image filtering, dense CRFS and bilateral neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4452–4461.
[172] J.T. Barron, B. Poole, The fast bilateral solver, in: Proceedings of the European Conference on Computer Vision, Springer, 2016, pp. 617–632.
[173] G. Bertasius, J. Shi, L. Torresani, Semantic segmentation with boundary neural fields, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3602–3610.
[174] L.-C. Chen, J.T. Barron, G. Papandreou, K. Murphy, A.L. Yuille, Semantic image segmentation with task-specific edge detection using CNNS and a discriminatively trained domain transform, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4545–4554.
[175] C. Peng, X. Zhang, G. Yu, G. Luo, J. Sun, Large kernel matters improve semantic segmentation by global convolutional network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 1743–1751.
[176] G. Bertasius, L. Torresani, X.Y. Stella, J. Shi, Convolutional random walk networks for semantic image segmentation., in: Proceedings of the CVPR, 2017, pp. 6137–6145.
[177] L. Lovász, et al., Random walks on graphs: a survey, Combinatorics Paul Erdos Eighty 2 (1) (1993) 1–46.
[178] A. Adams, J. Baek, M.A. Davis, Fast high-dimensional filtering using the permutohedral lattice, in: Proceedings of the Computer Graphics Forum, 29, Wiley Online Library, 2010, pp. 753–762.
[179] E.S. Gastal, M.M. Oliveira, Domain transform for edge-aware image and video processing, in: Proceedings of the ACM Transactions on Graphics (ToG), 30, ACM, 2011, p. 69.
[180] G.J. Brostow, J. Fauqueur, R. Cipolla, Semantic object classes in video: a high-definition ground truth database, Pattern Recognit. Lett. 30 (2) (2009) 88–97.
[181] T. Scharwächter, M. Enzweiler, U. Franke, S. Roth, Efficient multi-cue scene segmentation, in: Proceedings of the German Conference on Pattern Recognition, Springer, 2013, pp. 435–445.
[182] M. Cordts, M. Omran, S. Ramos, T. Scharwächter, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset, in: Proceedings of the CVPR Workshop on the Future of Datasets in Vision, 1, 2015, p. 3.
[183] G. Neuhold, T. Ollmann, S.R. Bulò, P. Kontschieder, The mapillary vistas dataset for semantic understanding of street scenes., in: Proceedings of the ICCV, 2017, pp. 5000–5009.
[184] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, R. Yang, The ApolloScape Dataset for Autonomous Driving (2018) arXiv preprint arXiv: 1803.06184.
[185] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? The Kitti vision benchmark suite, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, pp. 3354–3361.
[186] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, A.M. Lopez, The synthia dataset: a large collection of synthetic images for semantic segmentation of urban scenes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3234–3243.
[187] S.R. Richter, V. Vineet, S. Roth, V. Koltun, Playing for data: ground truth from computer games, in: Proceedings of the European Conference on Computer Vision, Springer, 2016, pp. 102–118.
[188] M. Everingham, S.A. Eslami, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes challenge: a retrospective, Int. J. Comput. Vis. 111 (1) (2015) 98–136.
[189] A. Krizhevsky, G. Hinton, Learning multiple layers of features from tiny images, Technical Report, Citeseer, 2009.
[190] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2009, pp. 248–255.
[191] F. Li, T. Kim, A. Humayun, D. Tsai, J.M. Rehg, Video segmentation by tracking many figure-ground segments, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 2192–2199.
[192] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, A. Yuille, The role of context for object detection and semantic segmentation in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 891–898.
[193] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: common objects in context, in: Proceedings of the European Conference on Computer Vision, Springer, 2014, pp. 740–755.
[194] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, A. Torralba, Semantic understanding of scenes through the ADE20K dataset, Int. J. Comput. Vis. (2018), doi:10.1007/s11263-018-1140-0.
[195] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, L. Van Gool, The 2017 davis challenge on video object segmentation (2017) arXiv preprint arXiv: 1704.00675.
[196] H. Caesar, J.R.R. Uijlings, V. Ferrari, COCO-stuff: thing and stuff classes in context, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018) 1209–1218.
[197] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor segmentation and support inference from RGBD images, in: Proceedings of the European Conference on Computer Vision, Springer, 2012, pp. 746–760.
[198] H.S. Koppula, A. Anand, T. Joachims, A. Saxena, Semantic labeling of 3d point clouds for indoor scenes, in: Proceedings of the Advances in Neural Information Processing Systems, 2011, pp. 244–252.
[199] A. Dai, A.X. Chang, M. Savva, M. Halber, T.A. Funkhouser, M. Nießner, Scannet: Richly-annotated 3d reconstructions of indoor scenes., in: Proceedings of the CVPR, 2, 2017, p. 10.
[200] I. Armeni, S. Sax, A.R. Zamir, S. Savarese, Joint 2D-3D-semantic data for indoor scene understanding, CoRR (2017). abs/1702.01105.
[201] J. Xiao, A. Owens, A. Torralba, Sun3d: A database of big spaces reconstructed using SFM and object labels, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1625–1632.
[202] S. Song, S.P. Lichtenberg, J. Xiao, Sun RGB-D: a RGB-D scene understanding benchmark suite, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 567–576.
[203] K. Lai, L. Bo, X. Ren, D. Fox, A large-scale hierarchical multi-view RGB-D object dataset, in: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2011, pp. 1817–1824.

[204] S.D. Jain, K. Grauman, Supervoxel-consistent foreground propagation in video, in: Proceedings of the European Conference on Computer Vision, Springer, 2014, pp. 656–671.
[205] J. Shotton, J. Winn, C. Rother, A. Criminisi, Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation, in: Proceedings of the European Conference on Computer Vision, Springer, 2006, pp. 1–15.
[206] M. Marszalek, C. Schmid, Accurate object localization with shape masks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2007, pp. 1–8.
[207] B.C. Russell, A. Torralba, K.P. Murphy, W.T. Freeman, Labelme: a database and web-based tool for image annotation, Int. J. Comput. Vis. 77 (1–3) (2008) 157–173.
[208] J. Tighe, S. Lazebnik, Superparsing: scalable nonparametric image parsing with superpixels, in: Proceedings of the European Conference on Computer Vision, Springer, 2010, pp. 352–365.
[209] S. Gould, R. Fulton, D. Koller, Decomposing a scene into geometric and semantically consistent regions, in: Proceedings of the IEEE 12th International Conference on Computer Vision, IEEE, 2009, pp. 1–8.
[210] B. Hariharan, P.A. Arbeláez, L.D. Bourdev, S. Maji, J. Malik, Semantic contours from inverse detectors, 2011 International Conference on Computer Vision (2011) 991–998.
[211] C. Liu, J. Yuen, A. Torralba, Nonparametric scene parsing via label transfer, IEEE Trans. Pattern Anal. Mach. Intell. 33 (12) (2011) 2368–2382.
[212] A. Valada, G.L. Oliveira, T. Brox, W. Burgard, Deep multispectral semantic scene understanding of forested environments using multimodal fusion, in: Proceedings of the International Symposium on Experimental Robotics, Springer, 2016, pp. 465–477.
[213] M.-Y. Liu, S. Lin, S. Ramalingam, O. Tuzel, Layered interpretation of street view images, CoRR (2015). abs/1506.04723.
[214] J. Krapac, I.K.S. Šegvic, Ladder-style densenets for semantic segmentation of large natural images, in: Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW), IEEE, 2017, pp. 238–245.
[215] W. Wang, U. Neumann, Depth-aware CNN for RGB-D segmentation, ECCV 2018, 2018.
[216] Z. Li, Y. Gan, X. Liang, Y. Yu, H. Cheng, L. Lin, LSTM-CF: unifying context modeling and fusion with LSTMS for RGB-D scene labeling, in: Proceedings of the European Conference on Computer Vision, Springer, 2016, pp. 541–557.
[217] Y. Wang, C. Xu, C. Xu, D. Tao, Beyond filters: Compact feature map for portable deep model, in: Proceedings of the International Conference on Machine Learning, 2017, pp. 3703–3711.
[218] Y.-D. Kim, E. Park, S. Yoo, T. Choi, L. Yang, D. Shin, Compression of deep convolutional neural networks for fast and low power mobile applications, CoRR (2015). abs/1511.06530.
[219] A. Holliday, M. Barekatain, J. Laurmaa, C. Kandaswamy, H. Prendinger, Speedup of deep learning ensembles for semantic segmentation using a model compression technique, Comput. Vis. Image Underst. 164 (2017) 16–26.
[220] J. Hoffman, D. Wang, F. Yu, T. Darrell, FCNs in the wild: Pixel-level adversarial and constraint-based adaptation, CoRR (2016). abs/1612.02649.
[221] Y. Zhang, P. David, B. Gong, Curriculum domain adaptation for semantic segmentation of urban scenes, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2, 2017, p. 6.
[222] S. Sankaranarayanan, Y. Balaji, A. Jain, S.N. Lim, R. Chellappa, Learning from synthetic data: addressing domain shift for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[223] N. Ma, X. Zhang, H.-T. Zheng, J. Sun, ShuffleNet V2: Practical guidelines for efficient CNN architecture design, ECCV, 2018.
[224] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, MobileNets: Efficient convolutional neural networks for mobile vision applications, CoRR (2017). abs/1704.04861.
[225] H. Li, A. Kadav, I. Durdanovic, H. Samet, H.P. Graf, Pruning filters for efficient convnets, CoRR (2016). abs/1608.08710.
[226] M. Jaderberg, A. Vedaldi, A. Zisserman, Speeding up convolutional neural networks with low rank expansions, CoRR (2014). abs/1405.3866.
[227] R. Zhang, G. Li, M. Li, L. Wang, Fusion of images and point clouds for the semantic segmentation of large-scale 3d scenes based on deep learning, ISPRS J. Photogramm. Remote Sens. 143 (2018) 85–96.
[228] M. Yousefhussien, D.J. Kelbe, E.J. Ientilucci, C. Salvaggio, A multi-scale fully convolutional network for semantic labeling of 3d point clouds, ISPRS J. Photogramm. Remote Sens. 143 (2018) 191–204.
[229] R.Q. Charles, H. Su, M. Kaichun, L.J. Guibas, Pointnet: deep learning on point sets for 3d classification and segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 77–85.
[230] Y. Li, J. Shi, D. Lin, Low-latency video semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5997–6005.



**Fahad Lateef** received his B.Sc. degree in electronic engineering from BUITEMS, Pakistan in 2009. Later in 2013, he received his M.Sc. degree in electronics from the Mid Sweden University, Sweden. He is currently doing his Ph.D. at LE2i-CNRS, University of Technology of Belfort-Montbéliard, France. He is a scholarship holder from Higher Education Commission of Pakistan. His research interests include machine learning, computer vision, artificial neural networks and semantic scene understanding.

**Yassine Ruichek** received the Ph.D. degree in control and computer engineering and the Habilitation à Diriger des Recherches (HDR) degree in physic science from the University of Lille, France, in 1997 and 2005, respectively. Since 2007, he has been a Full Professor with the University of Technology of Belfort-Montbéliard (France). His research interests are concerned with multisensory data based perception and localization, including computer vision, pattern recognition and classification, machine learning, data fusion, with applications to intelligent transportation systems and video surveillance.