SN Computer Science (2021) 2:397
https://doi.org/10.1007/s42979-021-00784-5

REVIEW ARTICLE

A Review on Progress in Semantic Image Segmentation and Its Application to Medical Images

Mithun Kumar Kar[1] · Malaya Kumar Nath[1] · Debanga Raj Neog[2]

Received: 14 July 2020 / Accepted: 19 July 2021 / Published online: 31 July 2021

Abstract

Semantic image segmentation is a popular image segmentation technique where each pixel in an image is labeled with an object class. This technique has become a vital part of image analysis nowadays as it facilitates the description, categorization, and visualization of the regions of interest in an image. The recent developments in computer vision algorithms and the increasing availability of large datasets have made semantic image segmentation very popular in the field of computer vision. Motivated by the human visual system which can identify objects in a complex scene very efficiently, researchers are interested in building a model that can semantically segment an image into meaningful object classes. This paper reviews deep learning-based semantic segmentation techniques that use deep neural network architectures for image segmentation of biomedical images. We have provided a discussion on the fundamental concepts related to deep learning methods used in semantic segmentation for the benefit of readers. The standard datasets and existing deep network architectures used in both medical and non-medical fields are discussed with their significance. Finally, this paper concludes by discussing the challenges and future research directions in the field of deep learning-based semantic segmentation for applications in the medical field.

Keywords Semantic segmentation · Deep learning · Automated medical image analysis · Convolution neural network · Deep neural network · Recurrent neural network

Introduction

Image segmentation plays an important role in computer vision applications as it influences all the critical tasks, such as image analysis, feature calculation, object detection and classification. Recently, with the advances in hardware technologies and development of neural network algorithms, emphasis is given on pixel level segmentation rather than localized segmentation of an image. This is called semantic image segmentation, where the different regions of an image can be clustered as different object classes. It exemplifies the process of combining each pixel of an image with a class label and gives multiple level of representation of the image by means of object classes. Nowadays scene parsing has become a fundamental research area in computer vision as the number of applications are on rise. Scene parsing is to analyze and segment an image into different image regions connected with semantic categories. It relies mostly on semantic segmentation [1–3]. For example, people may be interested in segmenting vehicles, persons, roads, and the sky in a traffic scene captured by a camera mounted on a vehicle to assist autonomous driving operations [2]. Some other important applications include detecting road signs [4], human machine interaction [5], virtual reality, and computational imaging [6]. These algorithms have found potential application in computer vision mainly due to its accuracy, which is achieved using emerging deep learning techniques like convolutional neural networks (CNN), deep neural networks (DNN) and recurrent neural networks (RNN) etc. Success of these techniques is attributed mainly to the increasing availability of datasets and increase in parallel computing

✉ Mithun Kumar Kar
mithunkar.iitg@gmail.com

Malaya Kumar Nath
malaya.nath@gmail.com

Debanga Raj Neog
debanga@alumni.ubc.ca

[1] Department of Electronics and Communication Engineering, National Institute of Technology Puducherry, Karaikal 609609, India

[2] Independent Researcher, Toronto, Canada

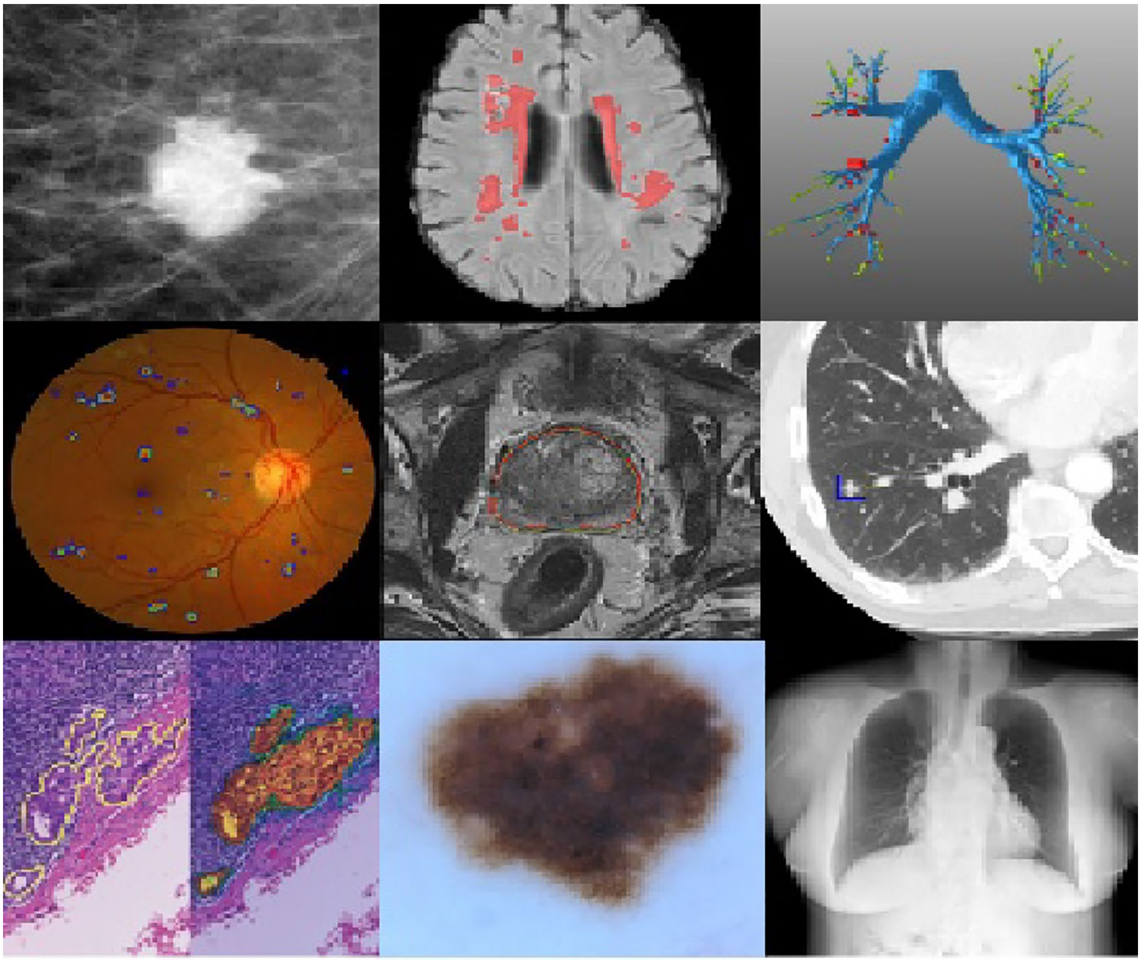


Fig. 1 Medical imaging applications in which deep learning has achieved state-of-the-art results. From left to right (top row): mammographic mass classification [7], segmentation of lesions in the brain [8], leak detection in airway tree segmentation [9]; from left to right (middle row): diabetic retinopathy classification [10], prostate segmentation [11], nodule classification [12]; from left to right (bottom row): breast cancer metastases detection in lymph nodes [13], human expert performance in skin lesion classification [14], and bone suppression in X-ray [15]. These images represented in this Figure have been extracted from [16]

power with widely available general-purpose graphics processing units (GPGPU). The deep learning methods ride into its popularity with it success in image classification, and its extensive research to perform pixel-wise image segmentation with more object classes in the scenes. This helps to provide well defined object boundaries. Both supervised and unsupervised machine learning methods have been successfully used for deep learning-based semantic segmentation tasks.

Recently, in the field of biomedical image processing, there is a huge increase in applications of image segmentation, recognition and registration techniques. The performance of image analysis by traditional methods, such as manual analysis of X-rays or CT scans, is somewhat restricted due to the limited experience of the analyzer, image complexity, non-similarity of interpretation and irregular anatomy between patients. Many of these limitations can be removed by the use of computer aided systems, and therefore, in the field of automated medical image analysis, applications of computer aided systems are on the rise.

These automated inspection methods have surpassed traditional methods by a large margin in terms of diagnostic measures. Some of the emerging applications where these techniques are showing promising results include glaucoma detection and blood vessels segmentation from fundus images [17, 18], brain tumor segmentation from MRI [19], segmentation of the pectoral muscle from breast MRI [20], segmentation of the coronary arteries in cardiac CT angiography, 3D segmentation in microscopic images [21], lesion segmentation [22], microscopy image analysis [23], colon crypts segmentation [24] to name a few. Figure 1 shows some of inferences of deep learning methods for classification and segmentation of biomedical images.

Researchers have provided review on semantic segmentation on natural and biomedical images. Many deep learning approaches for medical image segmentation have been introduced with different medical imaging modalities. Litigens et al. [16] reviewed major deep learning concepts related to medical image analysis. They reviewed deep CNN architectures for general classification, detection and segmentation of biomedical images. Thoma [25] reviewed semantic segmentation using traditional approaches. He focused on feature based approaches, unsupervised segmentation methods, random decision forest, conditional random fields and Markov random fields etc. Guo et al. [26] reviewed semantic segmentation of images using deep learning techniques by dividing the work into three categories: region-based, fully convolutional network (FCN) based, and weakly supervised segmentation methods. They discussed about major challenges and weaknesses of the deep learning methods based on data size, computational resources and accuracy of inferences. Liu et al. [27] presented a review on progress on semantic segmentation considering both traditional methods and deep learning techniques. They mainly focused on FCN, pyramid method in segmentation and multistage networks using convolutional neural networks (CNN). Goceri et al. [28] reviewed different challenges related to training of deep neural networks for segmentation of medical images. Taghanaki et al. [29] reviewed semantic segmentation of natural and medical images by categorizing the leading deep learning-based medical and non-medical image segmentation solutions into six main groups, like deep architectural, data synthesis, loss function, sequenced models, weakly supervised, and multi-task methods.

This review paper offers a complete overview of the deep learning based semantic segmentation techniques and their applications in biomedical imaging field. This also includes an overview of the state-of-the-art work, most recent datasets, details of the relevant deep learning techniques, potential research directions and open challenges in the field of biomedical imaging. The following contributions are highlighted in this review paper in comparison with the other existing surveys:

- A brief coverage of research contributions in the field of semantic segmentation of bio-medical images using DL techniques are made w.r.t modalities, types of organs and imaging applications. This paper discusses all important deep learning models used for semantic segmentation task.
- Popular deep architectures are discussed along with their applications and limitations.
- Different medical databases used for semantic segmentation of biomedical images with semantic segmentation ground truth have been explained. Along this, different open source software packages and libraries used for computation of deep learning algorithms have been presented for better understanding.
- A brief discussion about the deep learning architectures for semantic segmentation, including supervised and unsupervised learning models with their applications on medical imaging have been discussed.
- In the last, the paper highlights the important research directions and limitations of different model architectures w.r.t training, testing, hyper-parameter selection, modalities, and types of organs etc, for semantic segmentation in biomedical images.

The rest part of the paper is organized as follows: “Semantic Image Segmentation” summarizes the traditional segmentation algorithms and their characteristics. Semantic segmentation algorithms based on neural networks are discussed in “Neural Network-Based Methods for Semantic Image Segmentation”. Some popular deep network architectures are discussed in “Standard Deep Neural Network Architectures”. “Datasets” discusses about the available datasets and software. In “Deep Learning for Semantic Segmentation of Medical Images” the application of deep neural networks in medical imaging are described. Applications and challenges in semantic segmentation in biomedical field are discussed in “Discussion”. We conclude this paper in “Conclusion” with our comments on this review and future research directions.

Semantic Image Segmentation

Traditional segmentation methods focused on segmenting the region of interest while semantic segmentation segmented the different objects in an image to different classes. Based on the underlying technique of feature extraction, the semantic segmentation algorithms can be divided into two parts: traditional feature based classification methods and deep neural network-based methods. Traditional approaches use different featured-based classification methods, such as: region based segmentation [30], texton forest [31], random forest based classifiers [32], conditional random fields [33], and clustering techniques, where the features are hand crafted. On the other hand, neural network-based methods incorporate the domain knowledge available in a dataset through repeated spatial convolution operations to learn enhanced features for accurate inferences. Another way to categorize semantic segmentation algorithms is dividing into supervised and unsupervised segmentation methods. The supervised learning methods are influenced by supervision that uses intense domain knowledge or labeled data for separating the region of interest, whereas unsupervised learning develops perceptions right from the data itself, clusters the data and supports data driven decisions without any external bias.

Segmentation Architecture

Semantic segmentation architecture consists of classifiers which can classify the image into different semantic regions or assign each pixel a class label. Classifier-based methods depend on fixed size feature inputs and works on a predefined statistical or probabilistic model of the classifier. Statistical classifiers use supervised or unsupervised models for pixel classification, which directly depends on distribution of data. Probabilistic models used spatial probability distribution maps to tackle the variability of pixels. Markov random fields (MRF) and conditional random fields (CRF) are used for semantic segmentation, where the classifier learns the conditional distribution of the feature vectors for class labeling. Utilization of a CRF permits to include shape, texture, color, location, and edge cues in a single combined model. Another approach is a sliding window-based approach, where the trained classifier is fed with rectangular regions of the image and classifies the center pixel or a subset of the complete window. This type of approach is supported by neural network-based methods to handle a trained network as a convolution and apply the convolution on the complete image. Using deep neural networks and by increasing the layer of convolution it provides more satisfactory results than feature-based classifiers.

Traditional Methods for Semantic Image Segmentation

Traditional methods for semantic image segmentation rely on efficient feature detection and classification. In traditional methods, the main importance is given to feature detection or pixel wise classification or matching methods. Various hand crafted features are used for semantic segmentation, such as: pixel color [34], histogram of oriented gradients (HOG) [35], scale-invariant feature transform (SIFT) [36], local binary pattern (LBP) [37], sub-pixel corner [38] and features from accelerated segment test (FAST) [27]. Figures 2 and 3 represent semantic segmentation of biomedical images using traditional methods.

Chen et al. [39] proposed an intensity neighborhood-based supervised automated segmentation system for segmenting biomedical images. In training stage, the system received scaled, normalized input data and extracted significant pixels in neighborhood windows. Whereas in the testing stage, a voting procedure is used for predicting the unknown data with trained classifiers at different scales. Principal component analysis (PCA) is used to reduce the high dimensional complexity arising due to pixel windows. Brox et al. [40] proposed part-based poselet detectors, which use potential object contours and texture patches in the image for semantic object segmentation.

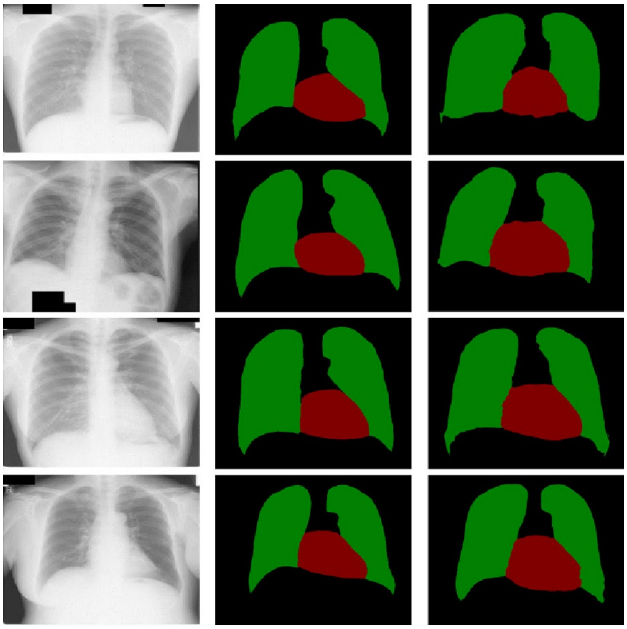


Fig. 2 Semantic segmentation on the JSRT dataset (red color represents the heart; green color represents the lungs). First column represents image, second column represents ground truth and third column represents prediction. This Figure has been extracted from [41]

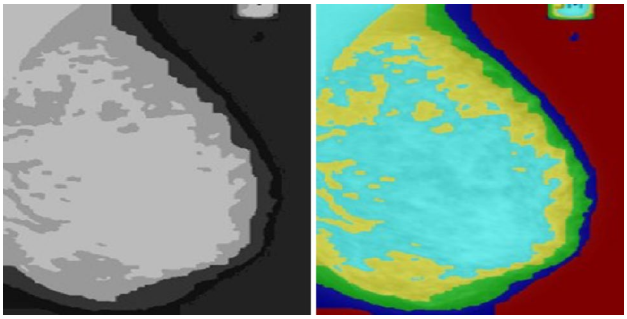


Fig. 3 Gray scale mammogram image and its semantic segmentation representation. This Figure is extracted from [42]

Adam et al. [43] used color cues for detection and classification of road signs using histograms of oriented gradient (HOG) descriptors and support vector machine (SVM) classifier. They applied the algorithm for Greek road signs detection and classifications. Also Dalal et al. [35] used HOG descriptor as an efficient feature for human detection. They used linear SVM classifier for detection.

Conditional random fields (CRF) are used in some work to exploit the spatial information for semantic image interpretation. Yang et al. [44] proposed a hierarchical conditional random field model for image classification by modeling spatial and hierarchical structures. They labeled the

Table 1 Traditional methods for segmentation of biomedical images with semantic segmentation ground truth

References	Imaging modalities	Organ of interest	Summary of techniques	Application
Thor et al. [48]	Mammogram	Breast	Watershed segmentation	1. Detected masses in digital mammograms
Yu-Len et al. [49]	Mammogram	Breast	Watershed segmentation	2. Brest tumor in 2D sonography
Gomez et al. [50]	Mammogram	Breast	Watershed segmentation	3. Breast nodules segmentation on ultrasonic images images
Nafiza et al. [42]	Mammogram	Breast	Graph cut techniques	Density based breast segmentation
Pan et al. [51]	MRI	Brain	Bayes-based region-growing algorithm	Segmenting brain MR images

image dataset with hierarchical CRF with the energy function given by

$$E(x|d) = \sum_{i \in V} E_1 x_i + \alpha \sum_{(i,j) \in N} E_2 x_i x_j + \beta \sum_{(i,k) \in H} E_3 x_i x_k, \qquad (1)$$

where α and β are weighting coefficients. x_i are the labels for each region i based on the image data d. E_1 represents unary potential, E_2 represents local pairwise potential which gives the relation between variables of neighboring regions within each scale and E_3 represents hierarchical pairwise potential, which represents relationships within regions of neighboring scales. They used 8-class eTRIMS dataset [45] which consists of 60 building facade images with 8 labelled classes. Shotton et al. [46] used appearance, shape and context information as a whole as textons (which jointly model shape and texture) for automatic visual recognition and semantic segmentation of photographs. They used a CRF model which integrates shape, texture, color, location and edge into a unified feature space. Dalal et al. [35] used histogram of oriented gradients as feature vectors and used SVM classifier for object/non-object classification. Also, Raviteja et al. [47] used Gaussian CRF model for semantic segmentation. Table 1 contains some important contributions towards semantic segmentation using traditional methods.

Neural Network-Based Methods for Semantic Image Segmentation

Recent developments in the field of neural networks have improved the state-of-the-art in semantic segmentation. Neural network-based classifiers use summation of weighted inputs with cascaded layers and apply activation functions to the weighted sum to obtain output. The general architecture consists of an input layer, several hidden layers, a fully connected layer, and an output layer. The mapping between the consecutive layers decides the architecture of the neural network. The network learns these parameters by updating the weights by minimizing an error function (cross entropy or mean squared error) [52].

In Fig. 4, X_i represents the layer of input neurons. The first hidden layer H_1 is represented by the function

$$y_j = f(Z_j), \qquad (2)$$

where $Z_j = \sum W_{ij} x_i$. Similarly, the second hidden layer H_2 is represented by the function

$$y_k = f(Z_k), \qquad (3)$$

where $Z_k = \sum W_{jk} x_j$. The final layer is represented by

$$y_l = f(Z_l), \qquad (4)$$

where $Z_l = \sum W_{kl} x_k$. The output of the final layer may pass through a threshold function to finally get the classified output.

Deep neural network-based methods permit efficient learning of features directly from the image data. The network learns the features that optimally represent the data for classification. The increasing use of convolutional neural networks (CNN), recursive neural networks (RNN), deep belief networks (DBN), and auto-encoders in image

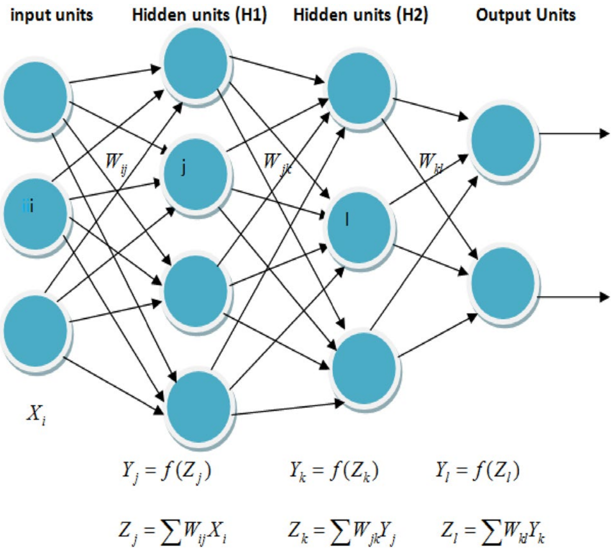


Fig. 4 Neural network architecture [52]

segmentation has enhanced the state-of-the-art in semantic segmentation of images [53, 54]. These deep learning algorithms use multiple layers of data abstraction and learn progressively high-level features to transform the input data into a suitable output form.

CNN in Semantic Segmentation

Convolutional neural networks (CNN) are artificial neural network architectures consist of several convolutional layers that allow spatial convolution of input image with different filter kernels to compute various feature maps [55]. It was first introduced by Kunihiko Fukushima [56]. In CNN, filters or kernels acting as weights slide across the total input image during convolution to create the next layer, and this new layer is called a feature map. The same filters can be used to repeat this operation to create new layers of feature maps. The input and output feature maps have different dimensions depending on the dimension of image channels, i.e. 1 or 3, respectively for grayscale and color images. Multiple filters can be applied across a set of input image slices where each filter will generate a distinctive output slice. These slices highlight the features detected by the filters. At each layer, the input image is convolved with a set of kernels or masks with added biases to generate a new feature map. If k number of kernels or masks are taken, each kernel can be represented by $W = \{w_1,w_2, w_3, \dots , w_k\}$ with added biases $B = \{b_1, b_2, b_3, \dots , b_k\}$. These feature maps are subjected to an element wise non-linear transformation, such as tanh, sigmoid, or rectified linear units (ReLUs)) and the process is repeated for every convolutional layer l [57]. Mathematically, it can be represented as

$$X_k^l = f(W_k^{l-1} * X^{l-1} + b_k^{l-1}), \tag{5}$$

where W_k represents the k^{th} kernel, X represents the input image or some predefined portion of the input image and b_k is the kth added bias. Figure 5 shows a basic convolution operation over an 3 × 2 matrix with a kernel size of 2 × 1.

After the convolution layer, pooling operation is performed which, downsamples each input feature map. The pooling process progressively reduces the dimensions of each input feature map while conserving the most significant features and avoids over-fitting [55]. Then outputs of each CNN layer are passed through non-linear activation functions, such as rectified linear units (ReLUs), which allow to represent composite non-linear mappings between the input image and the desired outputs. Figure 6 represents a basic classification model using CNN architecture.

Lo et al. [58] applied CNN for medical image analysis and hand-written digit recognition in LeNet [59]. But the popularity of CNN among the computer vision researchers started when a AlexNet proposed by Krizhevsky et al. [55] won ImageNet challenge in December 2012 by a huge margin. Later, more advanced CNN architectures have been proposed using related but deeper architectures. These architectures have shown outstanding performances on image segmentation, classification and detection tasks.

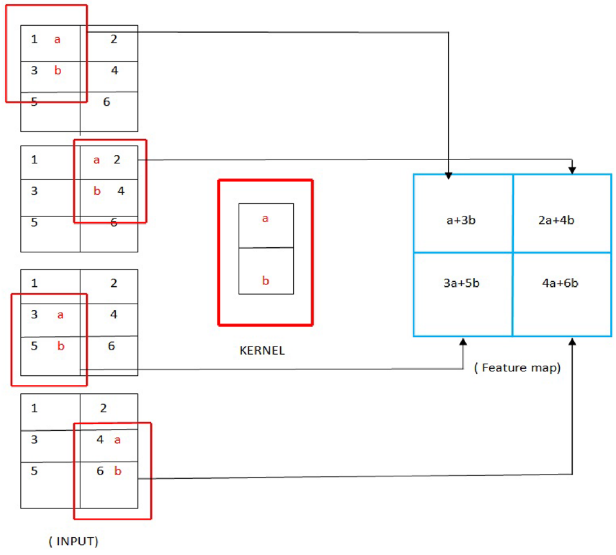


Fig. 5 Convolution operation

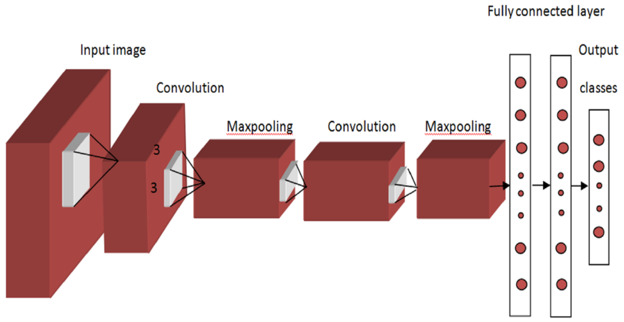


Fig. 6 CNN architecture [55]

Fully Convolutional Networks (FCN) in Semantic Segmentation

Fully convolutional network (FCN) proposed by Long et al. [60] has a fully connected convolutional layer in the last layer. A fully-connected layer can be taken as a special case of convolutional layer, with input volume of depth 1, with filter size same as the size of the input, and a total number of filters equal to the number of output neurons. FCN naturally works on an input of any size, and produces an output of consequent resampled spatial dimensions [60]. The different parts of a FCN (shown in Fig. 7) are convolution layers, pooling layers, activation functions, and softmax layers.

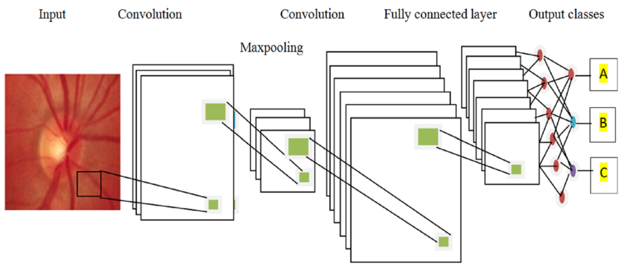


Fig. 7 FCNN architecture

Using pooling layers, convolutional layers can detect features at every level of the feature maps. The cumulative feature map size of deeper layers is larger than the ones at the beginning of the network. This permits them to capture more complex features from larger input regions. Generally the convolutional layers are used to extract features from the input in the form of feature maps. The features detected by the deepest layers are highly nonrepresentational. To solve this problem, one or more fully-connected convolutional layers are added after the last convolutional/pooling layer. The last fully convolutional layer (output) uses softmax to estimate the class probabilities of the input. Hence, this FCN layer can be treated as a translator between the network's representation and desired output [61].

Long et al. [60] used dense FCN for semantic segmentation, which combines dense downsampling layers and deconvolution layers (upsampling) to enhance spatial precision of the output. They used ImageNet large scale visual recognition challenge (ILSVRC) database [62] for dense prediction with upsampling and a pixelwise loss. They built a novel skip architecture which, combines different spatial information to refine predictions. They experimented the proposed algorithm on the PASCAL VOC 2011 segmentation challenge training dataset, NYUDv2 dataset [63] and SiftFlow dataset [64]. The same approach is adopted by Shelhamer et al. [65] which modified existing classification networks (AlexNet, VGG net, and GoogLeNet) into FCNs. They added fine-tuning to the segmentation task by defining a skip architecture that combines semantic information from deep layers that contain appearance information to produce detailed and high quality segmentation.

Atrous Convolution with CNN in Semantic Segmentation

Chen et al. [66] proposed a network architecture called DeepLabv3, which uses atrous convolution to extract dense features for semantic segmentation task. The network employs atrous convolution in cascade or in parallel form to capture multi-scale context by adopting multiple atrous rates, called atrous spatial pyramid pooling (ASPP). Here, the field of view of filters are enlarged effectively to include larger context without increasing the number of parameters or the amount of computation. In atrous convolution, the kernels are chosen with different dilation rates, which defines the spacing between the values in a kernel. Hence, a 3×3 kernel with dilation rate of two uses the same as 5×5, which covers a wider field of view at the same computational cost as a 3×3 kernel. Figure 8 shows schematic view of block representation of the atrous convolution.

There work has shown improved segmentation results due to the addition of encoder–decoder module [63], which shown in Fig. 9. In many notable work on image segmentation atrous convolution or dilated convolution is used for semantic segmentation to enhance the resolution of features as the quality of these features are often reduced due to repeated pooling operations or convolution striding in CNNs [53].

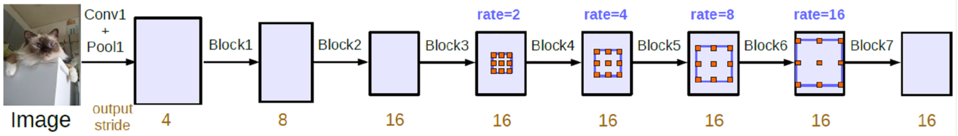
Recurrent Convolutional Neural Networks (RNN) in Semantic Segmentation

Recurrent neural networks (RNN) is used for processing sequential data with variable length. It used a recurrence relation between the current layer and previous layer using feedback loops. RNN can be defined as a recurrence relation

$$s_t = f(s_{t-1}, x_t), \tag{6}$$

where x_t is the network input at step t, y_t is the output of internal state at step t and s_{t-1} is the output of internal state at step $t - 1$. In a RNN, each state is dependent on all previous computations via this recurrence relation. Generally, RNN has three sets of weights:

Fig. 8 Deeper representation with atrous convolution. Figure has been extracted from [53]



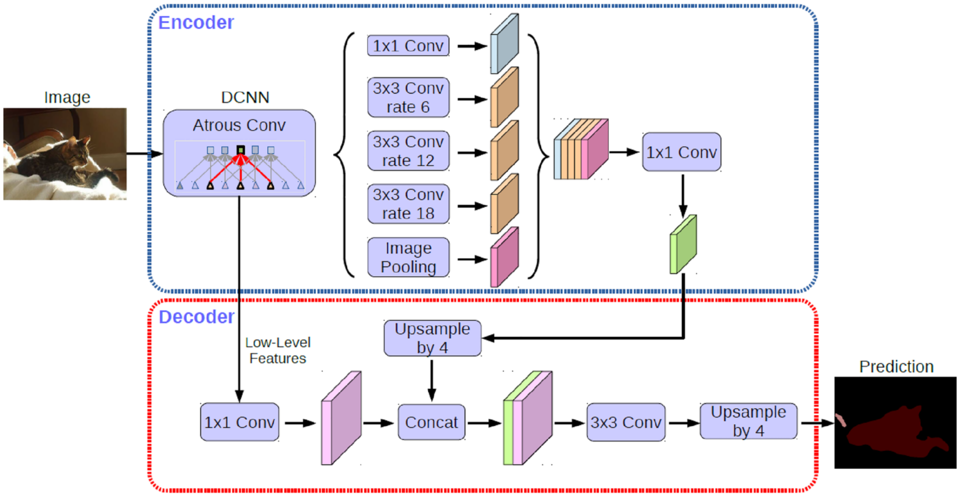


Fig. 9 DeepLabv3+ encoder-decoder structure. Figure has been extracted from [66]

1. U transforms the input x_t to the state s_t
2. W transforms the previous state s_{t-1} to the current state s_t
3. V maps the newly computed internal state s_t to the output y_t

The relation between the internal state and the network output is given as

$$s_t = f(s_{t-1} * Wx_t * U), \tag{7}$$

and

$$y_t = s_t * V, \tag{8}$$

where f is a non-linear activation function [67].

Pinheiro et al. [68] used recurrent CNN for scene parsing. They used a recurrent convolutional neural network on a large input image size context and trained the model in an end-to-end fashion over raw pixels using complex spatial dependencies with low inference cost. By increasing the context size with the built-in recurrence, the system itself determines and corrects its own errors. They used Stanford background dataset [69] and the SIFT flow dataset [64] for testing.

Recursive Context Propagation Networks (RCPN) for Semantic Segmentation

RCPN frames the problem of semantic segmentation as labeling of super-pixels [70] into desired semantic categories. It starts with the localization of semantically connected regions (super-pixels) followed by the extraction of visual features for each super-pixel. Multi-scale CNN [71] is used to extract per pixel features, which are averaged over super-pixels. Random binary parse trees are created with the adjacency information between super-pixels where leaf nodes correspond to initial super-pixels. Merging the nodes, a hierarchical graph structure is created, which is then passed through pre-designed modules to get the output labels. The final labels are decided through a voting procedure because each parse tree can give rise to different labels for the same super-pixel. Sharma et al. [72] used RCPN that employs contextual information of the whole image via random binary parse trees for improved feature representation of every super-pixel in the image. They compute bypass error paths in the computation graph of RCPN, which hamper contextual propagation. Hence they used pure-node RCPN and tree Markov random field-recursive context propagation network (MRF-RCPN) to minimize the bypass error.

Weakly-Supervised or Semi-Supervised Learning Models

In case of deep learning-based methods some weakly-supervised models have been proposed [73–75]. Training a deep neural network (DNN) requires a huge number of annotated segmentation ground truths to achieve good performance. Availability of consistent pixel-wise segmentation annotations are limited within a few popular datasets. Hence, it makes it difficult to use supervised DNNs in semantic segmentation tasks. In semi-supervised learning the unlabeled samples are used along with the labeled samples during training to improve the accuracy of the supervised learning with limited labeled samples [76]. Figure 10 shows the basic structure of semi-supervised learning.

Seunghoon et al. [73] proposed a decoupled DNN architecture using heterogeneous annotations, which is composed of two separately trained networks; one is for classification and the other one is for segmentation. The classification and segmentation networks are decoupled through bridging layers with class-specific activation maps, which deliver critical information from classification network to the segmentation

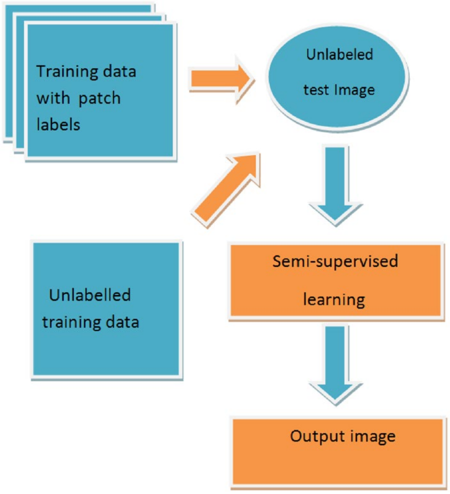


Fig. 10 Semi-supervised learning architecture. Figure has been reproduced from [77]

network. The object labels associated with an input image are recognized by the classification network while figure-ground segmentation of each identified label is obtained by segmentation network. The advantage of their model is that it uses pre-trained models for classification network and they train only segmentation network and bridging layers using a few strongly annotated data.

Kim et al. [54] proposed a framework for semantic segmentation using tied deconvolutional neural network with scale-invariant feature learning. In the proposed framework they have both convolutional layers and deconvolutional layers, where each deconvolution layer consists of unpooling and deconvolution using filter masks tied with that of the corresponding convolution layer. The restored features from all the deconvolution layers comprises a rich feature set and the feature maps with the uppermost abstraction level take out from the top most layers are used for reinforcement of final feature map. All the feature maps are concatenated across channel dimension, which covers all verities of features. Feature maps with the uppermost abstraction level take out from the top most convolution layer and the fine points of features are restored using deconvolution layers. Class-specific activation maps are generated using convolutional layers and are passed through softmax layers across channel dimensions and aggregated into a single vector to be compared with the image label vector.

Unsupervised Learning Models

Many authors have used unsupervised models in deep learning to overcome different challenges with deep neural networks, such as requirement of huge labeled datasets, overfitting in algorithms with supervision, and reduction of scalability of the target functions at hand. Layer wise unsupervised learning can be incorporated into deep neural architectures to improve accuracy where the data is not properly labeled, lack of annotated data, weak label annotations or when the amount of training data is less. Unsupervised learning algorithms can extract salient information about the input distribution, which reveals a representation that confines statistical regularities of the layers. It mainly helps to reduce the dependency on the changing gradient update direction given by a supervised criterion. Hence, unsupervised learning is an approach to naturally decompose the problem into sub-problems associated with different levels of abstraction [57]. They can be used to train deep neural networks for semantic image segmentation. These algorithms can be used as a part of supervised algorithms and trained to store information about the semantic classes. Unsupervised learning models can be used for semantic segmentation by utilizing some fixed models which are discussed below.

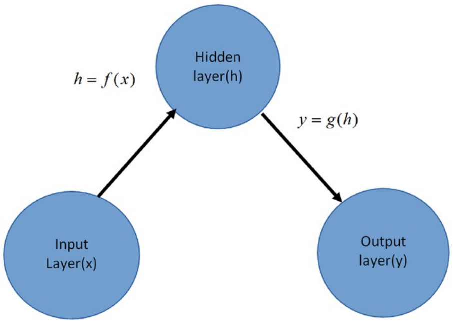


Fig. 11 Autoencoder maps an input layer x to an output layer y through a hidden layer h [77]

Autoencoders (AE) and Stacked Autoencoders (SAE)

Autoencoders (AE) [78] are unsupervised learning models comprise of a single-layer neural network. An autoencoder is trained to reproduce its inputs to the output layer through hidden layers as shown in Fig. 11. It mainly consists of two parts: first, an encoder which converts the input layer to a hidden layer, and second, a decoder that reconstructs the input from hidden layer. Originally they were used for feature learning with reduction of dimensionality, but now they are being used as latent variable models for regenerative modeling. The main idea behind copying the input layer to the output layer through hidden layers is to estimate the useful representative features. The input data can be projected on to a smaller dimensional subspace, which represents a dominant latent structure of the input and can be modeled to learn prominent features of the data distribution. It comes with a lot of variants, such as sparse autoencoder, convolutional autoencoder, variational autoencoder, contractive autoencoder, and denoising autoencoder. Also stacked autoencoders (SAE) are built by arranging autoencoders on top of each other and stacked into multiple layers, where the output of each layer is input to the successive layers. These layers are trained individually or in a greedy layer-wise fashion. Then, the full network is fine-tuned using supervised training to make predictions.

Contractive autoencoders use nonlinear loss functions and encourage the model to have properties, such as sparsity of the representation and noise robustness [79]. Chen et al. [80] used unsupervised learning using autoencoders for the classification of pulmonary nodules from lung CT images. They proposed a convolutional autoencoder neural network (CAENN) architecture for feature learning, which consists of an input layer, three convolution layers, three pooling layers and one fully connected layer. Denoising auto-encoders minimize the loss function of a copy of the input corrupted by some form of noise and it can minimize the reconstruction error. Gondara [81] used denoising autoencoders for efficient denoising of medical images.

Restricted Boltzmann Machines (RBM)

Boltzmann machines [82] are energy-based models, which are generally represented by the distribution function

$$p(x) = \exp(-E(x)), \tag{9}$$

where $E(x)$ is the energy function. The energy function of the Boltzmann machine is given by

$$E(x) = -x^{\mathsf{T}} wx. \tag{10}$$

Restricted Boltzmann machines (RBM) [82] are necessarily energy function-based undirected graphical models consisting of a single layer of latent variables used to learn the representation of input. They can be used to make deep graphical models that can learn the internal representation or the latent variables of the deep model by efficient interaction between the layers. A simple graphical representation is shown in Fig. 12.

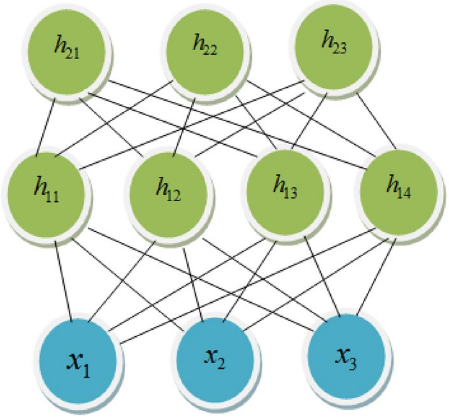


Fig. 12 Graphical model of RBM [77]

RBMs can be stacked to design deeper graphical models containing layers of observable variables and latent variables. The constrained connectivity between the layers makes it feasible to construct deeper models for efficient learning. RBMs have been extensively used in various parts of medical image analysis, such as image segmentation [83], feature learning [84], disease classification [85], mass detection in breast cancer [86], and brain lesion segmentation [84].

Deep Belief Networks (DBN)

DBN [87] is a generative hybrid graphical model having multiple hidden layers with no intra-layer connections between the hidden layers. Here, the probability distributions of all neurons can be copied to the next layer to learn the representation of the input. DBN can obtain both directed and undirected graph model and hence can be viewed as a mixture of unsupervised networks, such as RBMs and autoencoders. Without supervision in training stage, a DBN can learn the best features from the probability distribution of copied inputs in the hidden layer. Here, the connections between the top two layers are undirected, and therefore can be used for classification problems.

Transfer Learning

Transfer learning is generally used in the sense that the learned model parameters from a trained model can be transferred to train a new model. It is a machine learning method where prior knowledge of the model parameters of a working model are reprocessed as the starting point for training a new related model. It can be effective when a model is to be trained with small dataset or to be trained from scratch [88].

In deep learning scenario, a network trained on a large dataset can be used either as an initialization or a fixed feature extractor for a new model to be trained from a smaller dataset. It is already proven that to start with pre-trained weights is more helpful than random initialization of weights, even with large data sets [88, 90]. Also, it imposes constraints due to the size and similarity features between the datasets used in the trained model and to be used for training the new model.

Shie et al. [89] used transfer learning to overcome data scarcity and feature representation problem. They proposed a novel method for segmenting otitis media (OM) images, which is shown in Fig. 13. They learned a codebook in unsupervised manner by utilizing CNN with ImageNet dataset [91], then encoded OM images with the codebook to get weighting vector for each image. They applied these feature

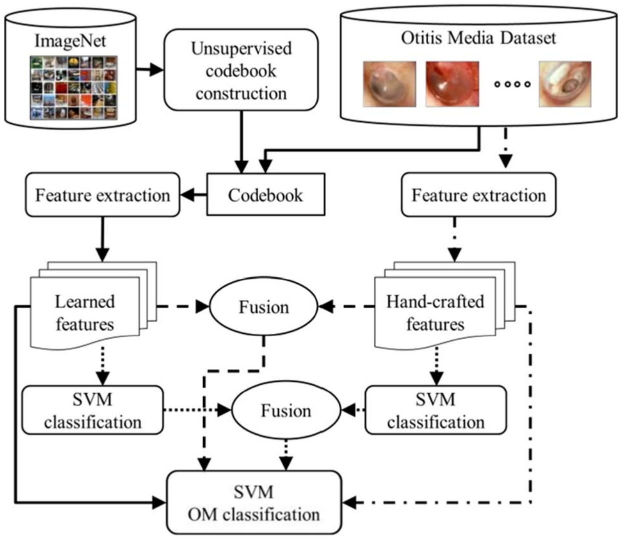


Fig. 13 Transfer learning model used for otitis media (OM) detection [89]

vectors to a supervised learning system using SVM to train an OM classifier with 1195 labeled OM instances. They achieved an accuracy of 88.5% for OM detection. Singh et al. [92] proposed a transfer learning based method for concept detection and modality classification.

Generative Adversarial Networks (GAN)

Generative adversarial network (GAN) is a deep model, first proposed by Goodfellow et al. [67] which is basically used to generate new replicas of data by learning the distribution of data. It consist of two neural network models called generator and discriminator. The generator captures the data distribution and generates unreal data and the discriminator tries to identify real datas from unreal datas. As a result of this competition, the discriminator and generator models are updated and the generator will update better-looking unreal data while the discriminator will become better at identifying them. The deep generative model generates the output from the input distribution which are looking same as input. The adversarial model is trained to optimally discriminate samples from the empirical data distribution and samples from the deep generative model. The basic block diagram of GAN is shown in Fig. 14 .

Luc et al. [93] trained a convolutional semantic segmentation network along with an adversarial network that discriminates segmentation maps coming either from the ground truth or from the segmentation network. It can detect and correct higher-order inconsistencies between ground truth segmentation maps and the ones produced by the segmentation net. They used the network on Stanford Background dataset and Pascal VOC 2012 dataset.

Instance-Aware Semantic Segmentation

Instance-aware semantic segmentation performs both classification and segmentation of object instances. It operates on region level and same pixel may have different semantics in different regions. Generally, segmentation is based on segment proposal and classification is based on region based methods.

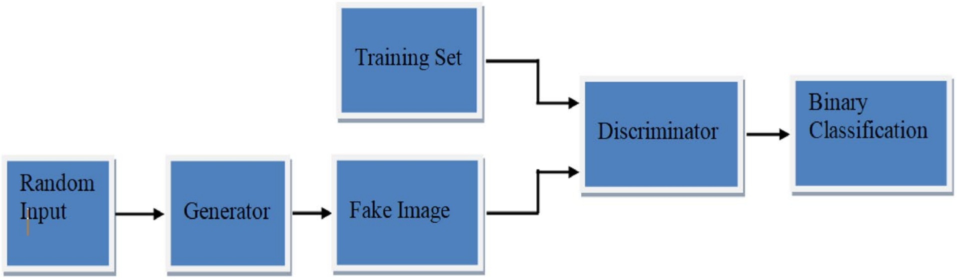
Li et al. [94] proposed a fully connected CNN for instance-aware semantic segmentation task (FCIS). FCIS uses rotation invariant property to perform both detection and segmentation. Authors created instant masks known as region-of-interest (ROI) from the FCN by region proposal network (RPN). This helps to produce pixel-wise score maps by assembling the operations in ROI. Detection and segmentation are the two tasks performed for each pixel in ROI. They trained two classifiers separately for mask predication and classification.

Multi task network cascades (MNC) is proposed by Dai et al. [95] for instance-aware semantic segmentation. The network consists of differentiating instances (represents by bounding boxes, which are class-agnostic), estimating masks (predicts pixel-level mask for each instance), and categorizing objects (predicts the categorize level). The network helps in sharing their convolutional features.

Standard Deep Neural Network Architectures

Several deep neural network architectures have been proposed in the last decade. Some of them gained popularity due to their enhanced performance in the fields, such as

Fig. 14 Basic block diagram of GAN



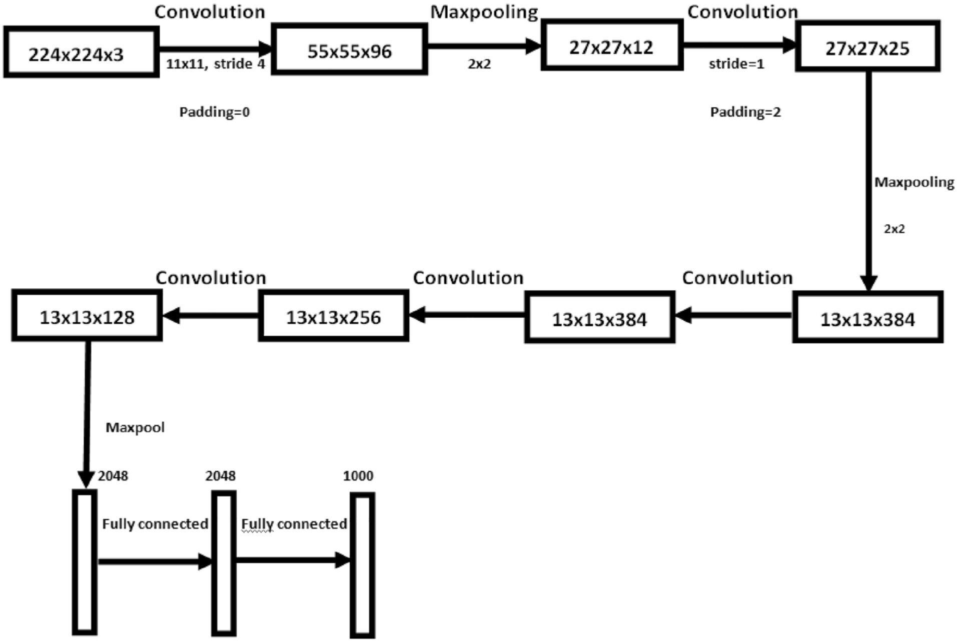


Fig. 15 AlexNet convolutional neural network architecture. Figure has been extracted from [55]

image classification, speech processing and robotics. These networks are becoming a standard choice for the researchers to solve novel challenges. Some of the important deep neural network architectures are reviewed below.

AlexNet

AlexNet was a revolutionary deep CNN architecture that won the ILSVRC-2012 challenge [55] with a accuracy of 84.6%, and it was a significant lead from the entries with traditional techniques which achieved a 73.8% accuracy in the same challenge. The architecture proposed by Krizhevsky et al. [55]. It contains eight learned layers (five convolutional layers with maxpooling layers and three fully-connected layers). ReLU non-linearity was applied to the output of every convolutional and fully-connected layers. A block diagram of the model is shown in Fig. 15. The model input image size was $224 \times 224 \times 3$ and classified to 1000 output classes.

VGG16

Visual geometry group (VGG16) is a CNN model proposed by Simonyan and Zisserman [96]. The architecture consists of a stack of 16 convolution layers followed by 3 fully connected layers with small receptive field of size 3×3. The model achieved 92.7% in top-five test accuracy in ImageNet large-scale visual recognition challenge (ILSVRC) 2014 with ImageNet dataset (consists of more than 10 million annotated images with 1000 classes). The block diagram of the mode is shown in Fig. 16.

GoogLeNet

GoogLeNet architecture (22 layer DNN) was introduced by Szegedy et al. [97], which won the ILSVRC-2014 challenge with 6.7% in top five error. It consists of stacked inception modules, which are convolutional neural networks with multiple receptive field sizes for convolution and pooling operation. They applied parallel filtering operations layer wise and concatenated all filter outputs together followed by 1×1 convolution operations to reduce dimensionality. The block diagram of an inception module is shown in Fig. 17 .

ResNet

ResNet architecture [98] was introduced by Microsoft corporation, which won ILSVRC-2016 challenge with an accuracy of 96.4%. It uses residual learning framework to train the dense layers of deep representations. The use of residual blocks with identity mapping helps in reducing the training errors due to large number of stacked layers. Figure 18 represents the building block of residual learning architecture.

ReNet

ReNet architecture was presented by Visin et al. [99] and used unidirectional RNNs. It uses four RNNs instead of CNNs, which sweep over the image patches in both horizontal and vertical directions. Composite feature maps are extracted from the intermediate hidden states by sweeping the RNNs both vertically and horizontally. Each subsequent

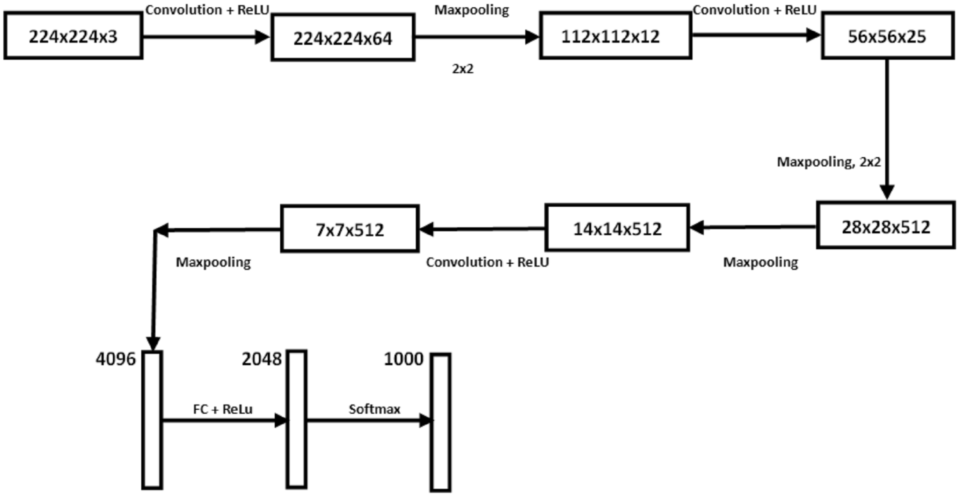


Fig. 16 VGG16 convolutional neural network architecture. Figure has been extracted from [96]

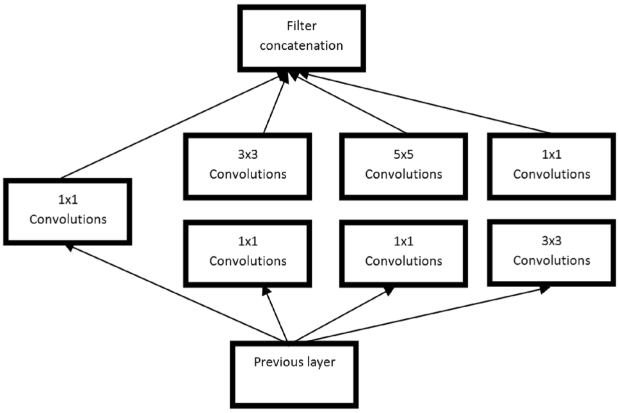


Fig. 17 Inception module with dimensionality reduction from the GoogLeNet architecture. Figure reproduced from [97]

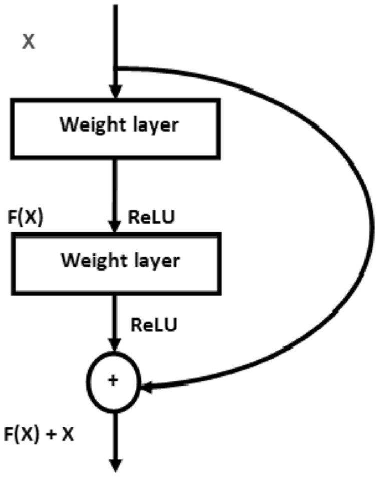


Fig. 18 Building block of residual learning. Figure extracted from from [98]

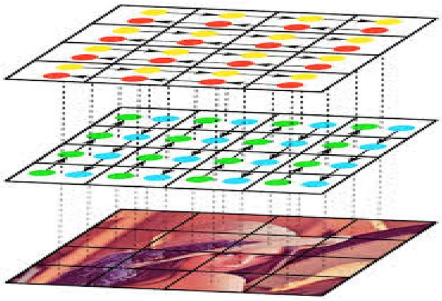


Fig. 19 One layer ReNet architecture. Figure extracted from [99]

layer operates on extracted representation from the previous layer, ensuring location specific operation. The output feature maps are stacked to create deeper architecture simultaneously capturing complex features. Figure 19 represents the basic structure of one layer ReNet architecture.

U-Net

U-Net is an encoder-decoder architecture first proposed by Ronneberger et al. [100], that have been used to segment biomedical images and a submission based on U-Net had won the international symposium on biomedical imaging (ISBI) cell tracking challenge in 2015. The network has a U-shaped architecture, which consists of two paths: one is a contracting path and the other one is a symmetric expanding path. Contracting path has general CNN structure consists of recurring layers of convolutions, followed by a rectified linear unit (ReLU) and a maxpooling operation. On the other hand, expanding path facilitates accurate localization of high resolution features. In contraction path multi-channel feature space is enhanced and spatial information is reduced, while expanding path uses a sequence of upsampling that allows the network to transmit perspective information to higher

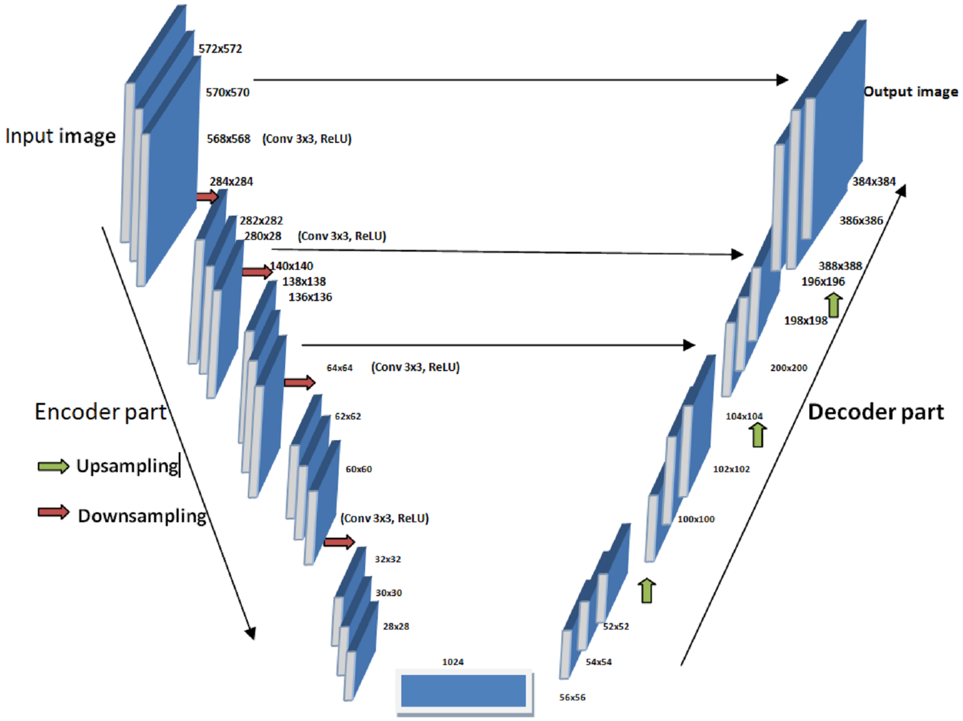


Fig. 20 U-Net architecture. Figure is extracted from [100]

resolution layers [100]. Figure 20 represents the basic structure of the proposed U-Net architecture. It has gain popularity for semantic segmentation of biomedical images.

SegNet

SegNet architecture is proposed by Badrinarayanan et al. [101] consists of an encoder-decoder neural network architecture used for semantic pixel-wise segmentation. The encoder part consists of 13 convolutional layers, which down-sample the input to low resolution feature maps preserving the high-level features. The decoder network is designed to up-sample the low resolution encoder feature maps to high resolution feature maps for pixel-wise classification. The decoder upsamples the low feature maps using pooling indices computed in the maxpooling steps corresponding to encoder which eliminates the need for up-sample learning. The upsampled maps are convolved with trainable kernels to yield dense feature maps and can be trained for pixel-wise classification. Figure 21 represents the basic structure of the proposed SegNet architecture.

PSPNet

The pyramid scene parsing network architecture (PSPNet) is proposed by Zhao et al. [102] for pixel-level scene parsing and it won ILSVRC 2016 challenge for scene parsing. It utilizes pyramid pooling module instead of global pooling to collect context information from the feature maps using CNNs. The pyramidal pooling uses four different pyramid scales to separate the feature maps into dissimilar regions with pooled representations. Bi-linear interpolation is used to upsample the low dimension feature maps to the appropriate size of original feature maps. Final prediction maps are generated by concatenating different size feature maps followed by a convolution layer. Figure 22 represents the overview of the proposed PSPNet architecture.

Datasets

In recent years, many large datasets are created by computer vision community with the emergence of deep learning models as they require large number of data samples to

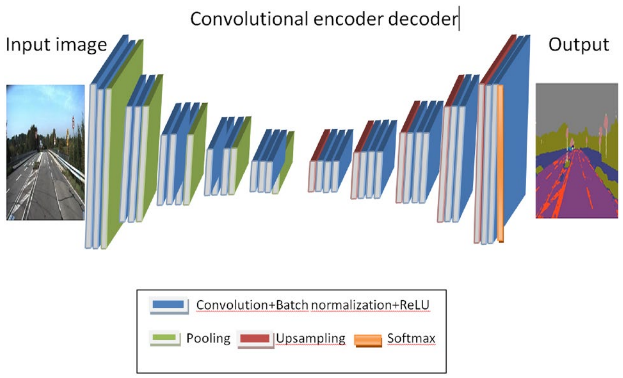

Convolutional encoder decoder
Input image
Output
Convolution+Batch normalization+ReLU
Pooling
Upsampling
Softmax

Fig. 21 SegNet architecture. Figure has been extracted from [101]

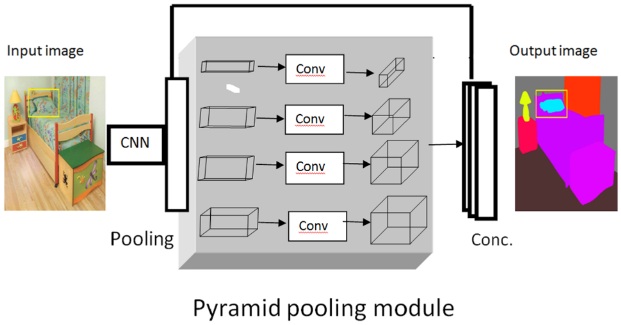


Fig. 22 PSPNet architecture. Figure has been extracted from [102]

train well. Here, some publicly available and widely used medical image datasets are mentioned. An overview of the datasets are given in Table 2.

Medical Image Databases for Semantic Segmentation

The DIARETDB1 [103] dataset contains 89 retinal fundus images, which can be used to detect diabetic retinopathy. The images are annotated with four classes: hard exudates, soft exudates, hemorrhages and red small dots.

Table 2 An overview of publicly available medical image datasets with semantic segmentation ground truth

Database	Modalities	Organs	Applications	Source
DIARETDB1	Fundus camera	Eye	Fundus images Diabetic retinopathy	Kalesnykiene et al. [103]
IDRiD				Prasanna et al. [104]
1. TCGA-LGG segmentation dataset	MRI images	Brain	Segmentation of cancer tissues	Setio et. al. [105]
2. BRATS 2015 dataset				Shaoguo et al. [106]
Open-CAS endoscopic	Endoscopic OCT	Pancreas	Medical instruments extraction	Maier et al. [107]
1. Warwick-QU	Microscopic image	Gland	1. Cancer gland segmentation	1. Coelho et al. [108]
2. Glas			2. Colorectal cancer detection	2. Gland segmentation in histology images challenge [109]
1. Fluo-N2DL-HeLa	Microscopic image	Cells	Microscopic cell segmentation	1. M. Maska et al. [110]
2. PhC-HeLa				2. Arteta et al. [111]
3. Hist-BM				3. Kainz et al. [112]
1. NIH database	X-ray	Chest	1. Chest X-ray	1. Xiaosong Wang et al. [113]
2. MIMIC-CXR	Radiographs		2. Chest radiographs	2. Johnson et al. [114]
3. JSTR database	Chest radiographs		3. Segmentation of the lung fields, the heart and the clavicles	3. Japanese journal of radiological technology [115]
4. SCR database	Chest radiographs		4. Segmentation of the lung fields, the heart and the clavicles	4. B. van et al. [116]
Colon Crypt DB	Colonoscopy videos	Colonic polyps	Segmentation of crypts in colon biopsies	Cohen et al. [24]
CQ-500	CT	Head	Head CT scan	Chilamkurthy et al. [117]
1. CATARACTS semantic segmentation dataset	Microscopic images	Tissue	1. Segmenting color images into body organs	1. Endoscopic vision challenge MICCAI 2020 [118]
2. Hamlyn centre laparoscopic/endoscopic video datasets	Microscopic images		2. In optical biopsy	2. M. Ye et al. [119]
1. Lung image database consortium image collection (LIDC-IDRI)	CT	Lungs	Lung cancer screening	1. https://public.cancerimagingarchive.net/ncia/login
2. Lung nodule analysis 2016 (LUNA16)	CT			2. https://public.cancerimagingarchive.net/ncia/login
3. Kaggles data science bowl, 2017 (DSB)	CT			3. www.kaggle.com/c/data-science-bowl-2017/data

IDRiD dataset [104] is used for detecting diabetic retinopathy (DR) and diabetic macular edema. It provides information regarding disease severity level of diabetic retinopathy and diabetic macular edema. The dataset consists of 81 color fundus images with pixel level annotation of abnormalities associated with DR, such as microaneurysms (MA), soft exudates (SE), hard exudates (HE) and hemorrhages. The images are stored in JPEG format with pixel resolution of 288 × 284 pixels.

The open-CAS endoscopic dataset [107] consists of 60 images taken from laparoscopic adrenalectomies and another 60 images taken from laparoscopic pancreatic resections. This dataset can be used for instrument segmentation in laparoscopic images. Semantic segmentation of these images help the surgeons to get sensory information about surgical procedures.

The Warwick-QU dataset consists of 165 images of colorectal cancer gland with pixel level annotation of 5 classes [108]. The classes may be divided in to healthy, adenomatous, moderately differentiated, moderately-to-poorly differentiated, and poorly differentiated. Semantic segmentation of these images help the medical practitioners to diagnose the cancer cells more accurately.

Fluo-N2DL-HeLa dataset contains frame sequences of cultured fluorescent HeLa cells which is used in ISBI cell tracking challenge [110]. In all frames, the ground truth contains markers for all 34060 cells and segmentation masks for 874 cells in four frames. This dataset is helpful for diagnosis of cell characteristics and semantic segmentation can be used to tackle the challenges like many cell clusters, frequent cell divisions, low contrast and variation in cell sizes.

PhC-HeLa dataset [111] contains 22 phase contrast microscopic images of cervical cancer colonies of HeLa cells. From these images, 2228 cells consist of cell markers with ground truth. This dataset is helpful for diagnosis of cell characteristics by semantically segment the cells with respect to ground truth so that high variation in cell shapes and sizes, missing cell boundaries, and high cell density regions can be accurately classified.

Hist-BM dataset [112] consists of 11 microscopic images containing hematoxylin and eosin of human bone marrow with ground truth consist of markers for all 4202 cell nuclei and unclear regions. This dataset can be used for semantic segmentation of cell nuclei and ambiguous regions and can be helpful for diagnosis of cell characteristics in microscopy image analysis.

The NIH Chest X-ray dataset [113] consists of 100,000 identified images of chest X-rays images with the text-mined fourteen disease image labels from 30,805 unique patients. The images are in PNG format. The data is provided by the NIH Clinical Centre and is available in NIH site. Fourteen common thoracic pathologies include atelectasis, consolidation, infiltration, pneumothorax, edema, emphysema, fibrosis, effusion, pneumonia, pleural thickening, cardiomegaly, nodule, mass and hernia. To create these labels, the authors used natural language processing to text-mine disease classifications from the associated radiological reports. The dataset can be used for semantic segmentation of common thorax diseases.

CQ-500 dataset [117] consist of 491 non-contrast head CT scans with 193,317 slices, provided by Centre for Advanced Research in Imaging, Neurosciences and Genomics, New Delhi, India. The dataset was used to detect intracranial hemorrhage (ICH) and its types (intraparenchymal hemorrhage (IPH), intraventricular hemorrhage (IVH), subdural hemorrhage (SDH), extradural hemorrhage (EDH) and subarachnoid hemorrhage (SAH), calvarial fractures, midline shift and mass effected) in head CT scans. The dataset can be used for semantic segmentation of intracranial hemorrhage and its types.

The MIMIC Chest X-ray (MIMIC-CXR) dataset [114] is a large publicly available dataset of chest radiographs with free-text radiology reports. The dataset consist of 377,110 chest X-rays associated with 227,827 imaging studies sourced from the Beth Israel Deaconess Medical Center between 2011 and 2016. Images are provided with 14 labels derived from two natural language processing tools applied to the corresponding free-text radiology reports. This dataset can be utilized for segmenting various pathologies related to lungs like enlarged cardiomediastinum, cardiomegaly, lung lesion, lung opacity, pneumonia and other abnormalities.

Gland Segmentation in Colon Histology Images Challenge Contest (GlaS) held at MICCAI 2015 [109]. The dataset used in this challenge consists of 165 images derived from 16 H and E stained histological sections of stage T3 or T4 colorectal adenocarcinoma. This dataset can be utilized for segmenting the extent of malignancy in histology images. The images in the dataset provides tissue architecture of two classes having benign and malignant histologic grades. The image can be semantically segment into challenging features like small glands, sub-mucosa layer, area with dense nuclei in mucosa layer and lumen of the gastrointestinal tract.

CATARACTS dataset (CaDIS) [118] is used for semantic segmentation of cataract surgery. It consist of 25 videos each having 30 frames per second for surgical procedure. Each video has a duration of 10 min and 56 s. The dataset consist of 29 surgical instrument classes, 4 anatomy classes and 3 miscellaneous classes and used for identification and localization of surgical instruments and anatomical structures through semantic segmentation.

The Lung Image Database Consortium image collection (LIDC-IDRI) consists of diagnostic and lung cancer screening thoracic computed tomography (CT) scans with marked-up annotated lesions. This dataset contains 1018 low-dose lung CTs taken from 1010 lung patients. It is used for evaluation of CAD methods for lung cancer detection and diagnosis. Semantic segmentation can be applied to the CT scans for segmenting lung cancer.

Hardware and Software

The rise of deep learning methods are principally due to the availability of large databases, wide availability of open source software packages, and increasing availability of consumer hardware, such as graphical processing units (GPUs) and GPU-computing libraries (e.g., CUDA, OpenCL).

GPUs are designed for faster and parallel processing of images in a frame buffer purported for output to a display device. Their parallel computing structure makes them more efficient for training massively parallelizable deep learning models.

The widely available open source software packages and libraries, developed based on the recent research in deep learning, is boosting the efficient use of deep learning methods in computer vision field. These libraries provide economical and efficient GPU implementations for the processing of large data with an acceptable processing time. Some popular packages are listed here.

1. MATLAB[1]: It offers specialised toolboxes for machine learning, neural networks and computer vision.
2. Caffe [62]: A deep learning framework developed by Berkeley AI Research (BAIR), provides C++ and python interfaces.
3. Tensorflow [120]: An open software math library, targeting mainly at implementation of deep learning models, provides C++ and python interfaces. It was developed at Google's AI research group.
4. Theano [121]: It provides a python interface designed to handle large neural network algorithms, developed by MILA lab in Montreal.
5. Torch [122]: It is a python based scientific computing package developed by Facebook's AI research group.
6. Keras [123]: An open source neural network library written in python, developed with a focus on enabling fast experimentation, supports both convolution based networks and recurrent networks.
7. MXNet [124]: An apache software foundation framework used to train and deploy deep neural networks.
8. Cognitive toolkit (CNTK) [125]: Frame work developed by Microsoft, which offers a python API over C++ code and operates under MIT license.

[1] The MathWorks, Inc., Natick, Massachusetts, United States.

Evaluation Measures for Semantic Segmentation

Different performance measures are adopted by computer vision researchers for semantic segmentation and scene parsing evaluations. Long et al. [60] used pixel accuracy, mean accuracy, mean intersection over union, and frequency weighted intersection over union for performance measures for image segmentation and defined as below:

$$\text{Pixel accuracy} = \frac{\sum_i n_{ii}}{\sum_i t_i}, \tag{11}$$

$$\text{Mean accuracy} = \frac{1}{n_{cl}} \sum_i \frac{n_{ii}}{t_i}, \tag{12}$$

Mean intersection over union

$$= \frac{1}{n_{cl}} \sum_i \frac{n_{ii}}{t_i + \sum_i n_{ji} - n_{ii}}, \tag{13}$$

Frequency weighted intersection over union

$$= \frac{1}{\sum_k t_k} \sum_i \frac{t_i n_{ii}}{t_i + \sum_i n_{ji} - n_{ii}}, \tag{14}$$

where n_{ij} is the number of pixels of class i predicted correctly to belong to class j where there are n_{cl} different classes, and $t_i = \sum_i n_{ij}$ is the total number of pixels of class i.

Sharma et al. [126] proposed recursive context propagation network (RCPN) for semantic segmentation where they took four standard evaluation metrics: per pixel accuracy (PPA), mean class accuracy (MCA), intersection over union (IoU) and time per image (TPI).

1. PPA is the ratio of the correctly classified pixels to the total pixels in the test image.
2. MCA is the mean of the category wise pixel accuracy.
3. IoU is the ratio of true positives to the sum of true positive, false positive and false negative, averaged over all classes.
4. TPI is the time required to label an image on GPU and CPU.

Roth et al. [127] computes dice similarity coefficient to measure the amount of agreement between two binary regions. To predict multiple classes for segmentation they have used a total loss function

$$K_{\text{total}} = \frac{1}{K} \sum_k w_k L_k, \tag{15}$$

where K is the number of classes (number of foreground classes and background) and w_k is a weight factor that can influence the contribution of each label class k. L_k is the loss function for each class k, which is given by

$$L_k = \frac{2\sum_i^N p_i r_i}{\sum_i^N p_i + \sum_i^N r_i} \tag{16}$$

where p_i represents the value of the probability map and r_i the corresponding ground truth at voxel i of total N voxels, present in the current image volume.

Deep Learning for Semantic Segmentation of Medical Images

Applications of deep learning to medical image analysis is now growing very rapidly. Due to its promising results, it is already being used in different fields, such as image segmentation [61], object detection [10, 128] and localization [22, 106]. The work has been categorized on the basis of imaging modalities, organs of interest and architecture. The mainly used architectures are fully convolutional networks (FCN), U-Net, RNN and GAN.

Fully Convolutional Networks in Semantic Segmentation of Medical Images

The fully convolutional network plays an important role in medical image segmentation due to its ability of dense prediction with pixel wise loss to predict different class at a time. This network has been used by the researchers for 2D and 3D images. The detail description about this network for 2D and 3D semantic image segmentation has been discussed in the following subsection

2D Images

Long et al. [60] used it for semantic segmentation of general scene images (ImageNet database) where they used last layer as fully convolutional layer instead of fully connected layer. Later skip architecture is added to it by Evan et al. [65] to fine tune the network. Cui et al. [106] segmented the intra-tumor structure of brain tumor using cascaded CNN with the MRI data. They proposed two subnetworks: tumor localization network (TLN) and intra-tumor classification network (ITCN). TLN is a fully convolutional network used to segment the tumor region from an MRI slice whereas ITCN is used to label the defined tumor region into multiple subregions. Akram et al. [23] applied FCN in microscopic image analysis for cell detection, segmentation and tracking. They proposed cell bounding boxes using a fully convolutional neural network (FCNN) and used a second CNN architecture to predict segmentation masks for each proposed bounding box. They used eight convolutional layers for detecting bounding boxes and used adaptive maxpooling inside the bounding box to extract fixed size features maps. By this network they were able to predict the regions that belonged to the cell to be segmented and accurately localize of cell boundaries. Tajbakhsh et al. [129] used fine tuned CNN instead of deep CNNs trained from scratch for polyp detection and pulmonary embolism detection using a free-response operating characteristic (FROC) analysis. Zhoua et al. [130] used weighted FCN with focal loss [131] to segment small objects as foreground. They emphasized on training wrongly-segmented pixels to decrease the number of false positives arise due to the uneven distribution of pixels in medical images.

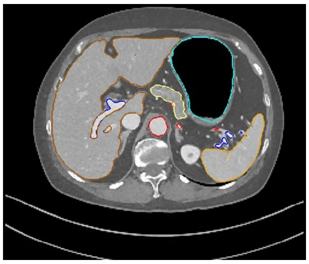


Fig. 23 Multi-organ segmentation in CT (axial). Figure reproduced from [77]

FCNN faces different challenges in the semantic segmentation of biomedical images. Due to the variable size of organs in biomedical images, the fixed perception field of FCNN, the same architecture is not able to produce satisfactory output on multiple organ segmentations. Many authors used multiscale FCN or cascaded FCN to overcome this. Xiangrong et al. [132] proposed a voxel-wise multiple-class classification scheme for automatically assigning labels to each pixel/voxel in a 2D/3D CT image.

3D Images

After successful application of FCN in image segmentation, researchers used 3D FCN, where both input image and kernels are in 3D form. Roth et al. [77] used 3D U-Net architecture [133] for multi organ semantic segmentation in CT images. It uses two paths in the network: analysis path and synthesis path. The analysis path is used for downsampling, which contains two convolutional layers followed by ReLU activation and a maxpooling layer. In the synthesis path, transposed convolutions are employed to convert the lower resolution feature maps within the network to the higher resolution space of the input images. It also utilizes skip architecture to provide higher-resolution features to the synthesis path. The final convolutional layer utilizes softmax activation function to compute a 3D probability map for

Table 3 Overview of papers using U-Net backbone for semantic segmentation tasks in biomedical images

Author	Imaging modalities	Organ of interest	Modified structure	Applications
Simindokht et al. [136]	Multi-modal	Lesion classification	mrU-Net	Skin lesion photos, lung CT, prostate magnetic resonance (MR) images
Ange et al. [137]	Thermography Electron microscopy (EM) Endoscopy images	Breast cancer, ventral nerve cord colonoscopy videos	DC U-Net	Semantic segmentation of organs
Yang Lei et al. [138]	Thorax CT images	AD/MCI classification	U-Net-GAN	Segment multiple OARs in thorax CT images
Zongwei et al. [139]	CT Slices	Lung nodules	U-Net++	Semantic segmentation of cell nuclei, brain tumor,liver and long nodule
Xieli et al. [140]	Glioma nuclei,	Cell nuclei	Dual U-Net	Segmentation of glioma nuclei

each of the target organs as the output of network. Figure 23 shows axial view of multi-organ segmentation of pancreas.

Ahn et al. [134] proposed a 3D convolutional neural network architecture called SqueezeNet3D for detecting lung cancer in the CT scans. In the Kaggle competition named Data Science Bowl 2017, the second winner Daniel Hammack [135] designed a 3D CNN for lung cancer classification. In this case the size of typical CT scans is about $512 \times 512 \times 400$ volumetric pixels. However, the region of interest is generally in the order of $1\ cm^3$. He used 3D CNN to predict nodule attributes to forecast a cancer diagnosis. He used 17 3D CNN models trained on LUNA16 dataset which consist of consist of 5 convolutional 3D blocks, followed by global maxpooling and a non-negative regression layer with a softplus activation. Kamnitsas et al. [21] proposed a 3D FCN for semantic segmentation of brain lesion. They employed a dual pathway architecture that processes the input images at multiple scales simultaneously. In post processing they used 3D fully connected conditional random field to remove false positives.

U-Net in Semantic Segmentation of Medical Images

But for biomedical image segmentation U-Net gets the high priority. Different modified structures are proposed based on U-Net backbone. It uses skip connections between the stages of network to regain the contextual information lost due to deep convolutional layers with pooling operations. Also it used concatenation of low-level features with high-level features for better learning representation. Many authors used modified U-Net architecture for multi-class segmentation of biomedical images [136–138]. Different methodologies based on U-Net are shown in Table 3.

3D U-Net in Semantic Segmentation of Medical Images

As majority of image modalities are in volumetric format, authors developed 3D U-Net based models to capture more affluent spatial information of volumetric images. Cicek et al. [133] proposed a 3D counterpart of U-Net architecture for volumetric semantic segmentation of xenopus kidney embryos at Nieuwkoop–Faber stage. The network consists of analysis and synthesis path as like 2D-U-Net but in the 3D form. The input to the network is a $132 \times 132 \times 116$ voxel tile of the image with three channels and the output in the final layer is $44 \times 44 \times 28$ voxels in x, y and z directions. The network output and the ground truth labels are compared using softmax with weighted cross-entropy loss. They used intersection over union (IoU) as performance matrices and infers the output with average IoU of 0.863.

The main issue with these 3D models is the memory limitations. As the voxels increase the parameters of the network, sophisticated and expensive hardware is required. Also the limited size of the voxels leaves constrains on resolution of the output. This can be overcome by dividing the input volume to multiple slices and used them for training and testing.

CNN with Residual Networks in Semantic Segmentation of Medical Images

CNN architecture with residual modules are proven to preserve more richer and discriminative feature information, lost by increasing the depth of the deep networks. Lequan et al. [141] used residual network with CNN for automated melanoma recognition from dermoscopy images. They proposed a fully convolutional residual network (FCRN) for accurate skin lesion segmentation. The network used 16 residual blocks in down-sampling path and three types of stride prediction map for upsampling. Hao et al. [142] used this residual learning concept on volumetric data and proposed VoxResNet architecture for volumetric image semantic segmentation of 3D brain MRI images. The VoxResNet architecture consist of stacked residual modules with a total of 25 volumetric convolutional layers and 4 deconvolutional layers. This can generate more representative features to deal with the large variations of brain tissues than fully connected layer. Also it has the capability to combine low-level image appearance features, implicit shape information, and high-level context together for semantic segmentation.

RNN in Semantic Segmentation of Medical Images

RNN architecture uses a recurrence relation between the current layer and previous layer using feedback loops, which empowered them to handle arbitrary input output length and to memorize the patterns from previous layer. RNN can be used with convolutional layers to capture the variations in pixel neighborhood. One of the successful variant of RNN is long short-term memory (LSTM) which is capable of learning long-term dependencies and can address the vanishing gradient problem generally occurs in simple RNN. In biomedical segmentation field many authors applied variants of LSTM such as bidirectional LSTM, convolutional LSTM (CLSTM), gated recurrent unit (GRU) [143] etc for segmenting multi-modal biomedical images.

Chen et al. [144] proposed a deep network for 3D image segmentation, based on a combination of a FCN and RNN called bidirectional LSTM (BDC-LSTM), which are responsible for utilizing both the intra-slice and inter-slice spatial dependences. The FCN take out and compress the hierarchy of intra-slice contexts into feature maps, and RNN (BDC-LSTM) extracts the 3D context from a sequence of preoccupied 2D contexts. Marijn et al. [145] proposed pyramidal multi-dimensional LSTM (PyraMiD-LSTM) network for segmentation of biological volumetric images, which employs six generalized convolutional LSTM (CLSTM) networks to develop the 3D context. This pyramidal structure is easier to parallelize and need less computations compared to multi-dimensional LSTM. Poudel et al. [146] proposed a recurrent FCN (RFCN) for automatic left ventricle segmentation from short-axis MR images of the left-ventricle using MICCAI 2009 challenge dataset [147].

Weakly-Supervised or Semi-supervised Learning Models in Semantic Segmentation of Medical Images

To overcome the constrains of supervised learning, mainly the shortage of pixel-level annotated databases in the medical imaging field with semantic segmentation ground truth, many authors developed weakly-supervised or semi-supervised learning models for semantic segmentation of biomedical images like autoencoders, restricted Boltzmann machines, deep belief networks etc.

Chen et al. [80] used unsupervised learning using autoencoders for the classification of pulmonary nodules from lung CT images. They proposed a convolutional autoencoder neural network (CAENN) architecture for feature learning, which consists of an input layer, three convolution layers, three pooling layers and one fully connected layer. Xinyang et al. [148] proposed a weakly-supervised nodule segmentation framework for segmentation of pulmonary nodules on lung computed tomography (CT) scans. The network consist of two stages like training stage and segmentation stage. In training stage, a CNN model is used to generate nodule activation maps (NAMs) for nodule localizations. In the second stage, coarse segmentation of nodule candidates are generated using multi-GAP CNN with residual connection. Several other works in the literature have used unsupervised models for semantic segmentation. Gu et al. [149] proposed a context encoder network (CE-Net) for 2D medical image segmentation. The network consists of three modules: a feature encoder module, a context extractor and a feature decoder module.They used a dense atrous convolution (DAC) block and a residual multi-kernel pooling (RMP) block. These two blocks are integrated with encoder–decoder structure to capture high-level features and preserve more spatial information. This method was applied on different medical images like fundus images for segmenting optic disc, blood vessels and retinal OCT layers. Chen et al. [80] proposed a convolutional autoencoder deep learning framework (CAENN) for classification of pulmonary nodules, which use unlabeled data for learning efficient features. They used 50000 unlabeled data of 64 × 64 patches, for unsupervised training and subsequently used 3662, 64 × 64 patches of labeled data for classification. Oktay et al. [150] proposed an anatomically constrained neural network model for pathological classification of cardiac MR and ultra sound images. It utilizes anatomical prior knowledge into CNNs and learns the representation of the underlying anatomy through a stacked convolutional autoencoder. Varghese et al. [151] used staked denoising autoencoders (SDAEs) for brain lesion detection, segmentation, and false-positive reduction. They used a large number of unlabeled data to train SDAEs and fine-tuned SDAEs with labeled patches drawn from a limited number of patients.

Transfer Learning in Semantic Segmentation of Medical Images

The idea before the transfer learning is to take the parameters of a trained model to initialize a new model, which can perform better than random initialization known as fine-tuning. For pixel-wise segmentation, the new network can be pretrained on a general scene database or biomedical image database. To capture the multi-modal perceptive of biomedical images it is better to fine tune the network with related domain. Transferring the weights from related domain is proven better than random initialization [152]. Tajbakhsh et al. [129] used transfer learning for fine tuning the network using AlexNet [55] and applied for polyp detection and pulmonary embolism detection.

Considering the weights from other domain, 25% higher sensitivity achieved compared to used CNN from scratch. Hariharan et al. [153] used CaffeNet [62] architecture built on ImageNet database for ultrasound kidney detection. They initialized the network with weights from CaffeNet parameters and the entire network weights were updated by training on kidney image samples. They achieved 4% increase in average dice overlap over the baseline method.

To overcome the difficulty of data acquisition and annotation in 3D medical imaging, Sihong et al. [154] used a heterogeneous 3D network called Med3D that contains a series of pretrained models. They converged eight medical databases to create 3DSeg-8 dataset that was used on pretrained Med3D network with ResNet [98] backbone. For lung segmentation task they achieved an accuracy of 93.82% which is much higher than baseline methods. These discussions favor the use of transfer learning approach over full training of a CNN from scratch. Lots of parameter are to be monitored when adapting transfer learning for biomedical images like modality domain, appearance and size of organs etc.

GAN in Semantic Segmentation of Medical Images

GAN are used to generate new duplicate data by learning the distribution of data. Yuan et al. [155] proposed a adversarial neural network, called SegAN for the task of medical image segmentation. They used a fully CNN as the segmentor to generate segmentation label maps, and proposed a novel adversarial critic network with a multi-scale $L1$ loss function to force the critic and segmentor to learn both global and local features that capture long and short range spatial relationships between pixels. Rezaei et al. [156] proposed recurrent GNN (RGNN) for medical image semantic segmentation. They used U-Net with skip connection as recurrent generator G to produce segmentation maps and recurrent discriminator R as classifier. Bidirectional LSTM was used as the recurrent architecture selected for both discriminator and generator models. They used mixed categorical loss function for training procedure of the semantic segmentation task, which helps to tackle the effect of imbalanced training data problem (Tables 4, 5, 6).

Discussion

As semantic segmentation is well suited to handle the variations of distribution of pixels into semantic classes, it is increasingly being used for related image analysis applications. In medical image segmentation, deep learning-based techniques are overpowering the traditional feature-based methods in usage. Therefore, in this paper attention is given on reviewing image segmentation techniques based on deep neural networks. These techniques allow automatic learning of the best features unlike traditional methods, where features are handcrafted, like pixel color [34], histogram of oriented gradients (HOG) [35], scale-invariant feature transform (SIFT) [36], local binary pattern (LBP) [37], sub-pixel corner [38]. Furthermore, accurate segmentation leads to better object detection and classification.

However, deep learning-based methods come with lots of challenges. Some of the related issues are discussed in the following sub sections.

Challenges with the Existing Databases

Generally, all deep learning algorithms work on annotated datasets along with ground truths. Datasets are fine tuned for specific context and how to achieve generality is still unclear. Also, collecting labeled datasets with dense annotations is expensive and time consuming. Another important issue is the requirement of expensive and high performing computing hardware. In particular, to reduce the training time, large number of graphical processing units are required while training neural networks with dense convolution layers. For medical images till now there is a lack of databases having large number of annotated ground truths. The number of increased images will improve the performance of deep neural networks for automated multi-organ segmentation in medical imaging. This hampers the growth of supervised deep learning algorithms to be applied to the medical databases compared to general databases with semantic ground truth. Generally data augmentation techniques are incorporated to increase the size of the databases. These augmentation techniques mainly consist of applications of affine transformations (rotation, flipping and scaling etc). It is already experimented that these augmentation techniques are able to increase the performance up to several percentage [179]. Recently, many authors proposed GAN for increasing the number of images in the datasets [180].

Challenges in the Deep Learning Architectures

Different deep learning structures have their own characteristics and mainly database dependable. For general scene parsing starting with AlexNet, till now different backbone structures are proposed (VGG16, GoogLeNet, ResNet, ReNet, SegNet etc). The AlexNet has the ability for object detection and classification in large datasets like ImageNet. However any small change in the convolutional layer will significantly degrade AlexNet's performance. VGG16 was one of the best performing architecture in ILSVRC challenge

Table 4 Overview of papers using deep learning for various image analysis and semantic segmentation tasks

Author	Imaging modalities	Organ of interest	Method used	Applications
Feng et al. [157]	Chest radiographs	Heart and lungs	ReNet and LSTM.	Semantic segmentation on the JSRT dataset
Roth et al. [77]	CT	Pancreas	CNN	Orthogonal patches from superpixel regions are fed into CNNs in three different ways
Roth et al. [127]	CT	Abdominal multi-organ segmentation	3D CNN	Volumetric pancreas segmentation
Cai et al. [158]	CT	Pancreas	CNN+CRF	2 CNNs detect inside and boundary of organ, initializes conditional random field.
Farag et al. [41]	CT	Pancreas	CNN	Approach with elements similar to Roth et al.
Chilamkurthy et al. [117]	CT	Head	CNN	Detection of critical findings in head CT scans
Thong [159]	CT	Kidney	CNN	Kidney segmentation in contrast-enhanced CT scans
Charbonnier [9]	Chest CT	Airway segmentation	CNN	Leak detection in CT scans
Lessmann et al. [160]	CT	Chest	CNN	Automatic coronary calcium scoring
Kai Hu et al. [17]	Fundus image	Eye	CNN	Retinal vessel segmentation from color fundus images using multi-scale CNN
Xiangyu et al. [18]	Fundus image	Eye	CNN	Glaucoma detection based on deep CNN
Juan et al. [161]	Fundus image	Eye	CNN	Glaucoma classification using CNN and transfer learning
Zhixi et al. [162]	Fundus image	Eye	CNN	Glaucoma classification using CNN and transfer learning
Raghavendra et al. [163]	Fundus image	Eye	CNN	Glaucoma classification based on CNN and Transfer learning in color fundus images
Grinsven et al. [10]	Fundus image	Eye	CNN	Hemorrhage detection in color fundus images
Shaoguo et al. [106]	MRI	Brain tumor segmentation	CNN	Semantic segmentation of brain gliomas
Pereira et al. [84]	MRI	Brain	CNN	Brain tumor segmentation
Havaei et al. [19]	MRI	Brain	CNN	Brain tumor segmentation
Karimi et al. [11]	MRI	Prostate	CNN	Prostate segmentation
Dong et al. [164]	MRI	Brain	CNN	Brain tumor detection and segmentation
Menze et al. [2]	MRI	Brain	CNN	Multimodal brain tumor segmentation

2014. It is very slow to train. The size of VGG16 trained ImageNet weights is 528 MB. So, it acquires more disk space and bandwidth that makes it incompetent for large databases. SegNet is a deep encoder-decoder architecture used for semantic segmentation for general scene datasets and applications like autonomous driving, scene understanding, etc. But, due to maxpooling and upsampling in encoder-decoder architecture there is a significant reduction in feature map resolution, which infers low output resolution. To overcome this issue Zhao et al. [102] proposed pyramidal pooling module instead of global pooling to preserve the context information using CNN.

But, for biomedical image segmentation UNet gets the high priority. Different modified structures are proposed based on UNet backbone. It uses skip connections between the stages of network to regain the contextual information, which is lost due to deep convolutional layers with pooling operations. Also it used concatenation of low-level features with high-level features for better learning representation. Many authors used modified UNet

Table 5 Overview of papers using deep learning for various image analysis and semantic segmentation tasks

Author	Imaging modalities	Organ of interest	Method used	Applications
Dvorak et al. [165]	MRI	Brain	Structured prediction+CNN	Multimodal brain tumor segmentation
Alansary et al. [166]	MRI	Uterus	CNN	Segmentation of the human placenta from motion corrupted MRI
Huynh et al. [20]	MRI	Breast	CNN and transfer learning	Mammographic tumor classification
Smistad et al. [128]	Ultrasound	Femoral region of both leg	CNN	Vessel detection in ultrasound images
Chen H et al. [39]	Ultrasound	Heart	CNN	Anatomical structure detection and segmentation from ultrasound images
Gao et al. [167]	Ultrasound	Obstetric ultrasound videos	CNN	Anatomical structure detection and segmentation
Li et al. [168]	Microscopic image	Breast cancer	CNN +CRF	Cancer metastasis structure detection
Akram et al. [23]	Microscopy image	Cell	CNN	Cell segmentation
Ronneberger et al. [61]	Microscopy image	Cell	CNN	Cell segmentation
Oren et al. [169]	Microscopy image	Mammalian cell	CNN with MIL	Instance learning
Ariel et al. [170]	Microscopy image	Cell	CNN with MIL	Multiple sclerosis (MS) lesions segmentation
Esteva et al. [14]	Microscopy image	Skin cell	DNN	Classification of skin cancer cell
Fotin et al. [171]	Microscopy image	Breast tissue	DNN	Detection of soft tissue densities from digital breast tomosynthesis
Giang et al. [12]	Microscopy image	Lungs	CNN	Lung nodule classification
Kooi et al. [7]	Mammogram	Breast	CNN	Mammographic lesions detection
Kamnitsas et al. [21]	MRI	Brain	3D CNN and fully connected CRF	Brain lesion segmentation
Ramaswamy et al. [172]	CT scan	Lungs	CNN	Pulmonary nodule classification
Kawahara et al. [22]	Microscopic image	Skin	CNN	Detection of skin lesions
Hammack et al. [135]	CT scan	Lung cell	CNN	Lung cancer diagnoses

architecture for multi-class segmentation of biomedical images [136–138]. Different methodologies based on UNet are shown in Table 3. Despite their success, these models suffer depth optimization and unnecessary complexity due to fusion of skip connections. These limitations may be overcome by modifying the structure and redesigning the skip connections.

Challenges with Training, Testing and Hyper-parameter Selection

Training a deep model with semantic segmentation ground truth is tricky as it involves optimization of different parameters regarding training, testing and hyper-parameter selection. For training a network with a large database with a large number of cascaded stages require a large training time. Reduce the training time and faster convergence are the issues in CNN. These can be achieved by reducing the dimensionality of the parameters. Generally pooling layers are used to reduce the dimension of the feature maps. Pooling variants are maxpooling, average pooling and adaptive pooling. Also many authors used convolution with variable stride to lighten the network or minimizing the parametres. But it has the effect of information loss. Batch normalization is used to reduce internal co-variate shift and provides faster convergence. It is performed through subtracting mean from the mini batch output and normalized by standard deviation of the mini batch. It is also known as an effective key for faster convergence. Batch normalization is a more preferred approach to improve the network convergence as it is not reported to have any negative effects on the performance, while the pooling and down-sampling techniques have came out with loosing beneficial information.

Another two important parameters are over-fitting and vanishing gradient problem. Over-fitting occurs when a model can learn well on training data by capturing the patterns and regularities in the training set with higher

Table 6 Overview of papers using deep learning for various image analysis and semantic segmentation tasks using unsupervised models

Author	Imaging modalities	Organ of interest	method used	Applications
Zhu et al. [31]	MRI	Lesion classification	SAE	Hierarchical classification to detect prostate cancer
Avendi et al. [173]	MRI	Segmentation	SAE	Segmentation of right ventricle in cardiac MRI
Suk et al. [27]	MRI	AD/MCI classification	SAE	SAE accompanied by supervised fine tuning for AD/MCI classification
Guo et al. [174]	MRI	Hippocampus segmentation	SAE	Representation learning and measure target/atlas patch
Mansoor et al. [175]	MRI	Visual pathway segmentation	SAE	To learn appearance features to steer the shape model for segmentation
Su et al. [176]	Microscopic image	Cell segmentation	SDAE	Structured labels for cell segmentation
Cheng et al. [34]	Ultrasound	Breast	SDAE	Stacked denoising AE for diagnosis of breast nodules and lesions
Cai et al. [177]	CT, MRI	Vertebrae localization	RBM	RBMs to locate the exact position of the vertebrae
Cao et al. [86]	Mammography image	Mass detection in breast cancer	RBM	Cell segmentation
Brosch et al. [83]	3D MRI	Multiple sclerosis segmentation	RBM	Uses 3D MR images of multiple sclerosis (MS) for MS segmentation
Pereira et al. [84]	MRI image	Brain lesion segmentation	RBM	RBM is used for feature learning
Azizi et al. [178]	Ultrasound image	Lesion classification	DBN	Training DBN to extract features for lesion classification

accuracy, but performs inadequately on unseen data with lower accuracy. Over-fitting can be reduced by increasing the size of the training database, hence it is an important issue for medical image databases. Different augmentation techniques are used to increase the size of the dataset by also maintaining the variability. Another regularization technique called dropout [181], is used by different authors for reducing the over-fitting at the time of training. Here, randomly selected neurons are dropped out from the activation or temporally removed on the forward pass during training. This will help the network to learn independent internal representations, which helps to reduce over-fitting.

Another problem faced by deep neural networks with gradient based optimization, is vanishing gradient problem. This occurs due to increase in hidden layers and the gradient becomes zero before the convergence. Hence, the error gradient cannot be efficiently back propagated to lower layers, inferring unsatisfactory result. One solution is to increase the number of training data so that the gradient vector spans the total epochs of the training. Many methods have been proposed, such as alternate weight initialization schemes [182], unsupervised pre-training [183], guided layer-wise training [184] and variations on gradient descent. Authors used ReLU, which prevents the gradient to diminish.

Challenges with Organ Appearance

Both in the 2D and 3D biomedical images, the diverAse and overlapped surfacing of the organ provides a big challenge to the researchers in segmentation field. The varying size of the organs and the indistinct boundary between target organ and its neighboring tissues in the imaging, give a ill-posed problem in the segmentation field. However, deep architectures with multitask learning network can address this issue [140]. Applying weighted loss function with a larger weight allocated to the separating background labels between touching organs is another successful approach for touching objects of the same class.

Challenges with Image Dimension

In dealing with 3D biomedical images, the training of the volumetric data faces more challenges than 2D images. Due to limited amount of training data, large number of parameters and high memory requirement, make the training much more expensive and time consuming for producing satisfactory inferences. It is not always possible to get fully annotating data for 3D images. Hence sparsely annotated data are used, which makes the inference with low accuracy. Weighted loss functions can be used to handle these unbalance sparsely annotated volume data [133].

Conclusion

Nowadays, semantic image segmentation has become an essential task in many applications in the field of computer vision and machine learning, where multiclass segmentation task is of importance. In the field of medical imaging the application of semantic segmentation is constantly growing. This paper covered fundamentals about deep learning techniques, databases used in both medical and non-medical field and their structures. It over viewed recent progresses in semantic segmentation in biomedical field, especially deep neural network-based semantic segmentation techniques. Deep learning-based semantic image segmentation techniques are more accurate and with the increasing availability of datasets and graphical processing units the inference time is decreasing with time. In the field of medical imaging, the main challenges remain due to the variability of patient data and lack of large datasets.

Acknowledgements This work is carried out at NIT Puducherry, Karaikal, India.

Declarations

Conflict of Interest There is no conflicts of interest.

Ethical Standards This paper satisfies and fulfills the compliance with ethical standards. The research doesn't involve human participants and/or animals.

References

1. Fritsch Kuehnl J, Geiger A. A newperformance measure and evaluation benchmark for road detection algorithms. In: Proc. IEEE Int'l Conf. Intelligent Transportation Systems (ITSC). 2013;9-18.
2. Menze M, Geiger A. Object scene flow for autonomous vehicles. Computer Vision and Pattern Recognition: Proc. IEEE Int'l Conf. 2015:1-11.
3. Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S, Schiele B. The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016;3213-3223.
4. Huang C, Davis L, Townshend J. An assessment of support vector machines for land cover classification. In: Proc. IEEE Int'l Journal Remote Sensing; 2002;23(4):725-749.
5. Oberweger M, Wohlhart P, Lepetit V. "Hands deep in deep learning for hand pose estimation. 2015;1-10. arXiv:1502.06807
6. Yoon Y, Jeon HG, Yoo D, Lee JY, Kweon IS. Learning a deep convolutional network for light-field image super resolution. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. 2015;24-32.
7. Kooi T, Litjens G, van Ginneken B, Gubern-Merida A, Sanchez CI, Mann R, den Heeten A, Karssemeijer N. Large scale deep learning for computer aided detection of mammographic lesions. Med Image Anal. 2016;35:302–312.
8. Ghafoorian M, Karssemeijer N, Heskes T, van Uden IWM, de Leeuw F, Marchiori E, van Ginneken B, Platel B. Non-uniform patch sampling with deep convolutional neural networks for white matter hyper intensity segmentation. IEEE Int Symp Biomedical Imaging. 2016;1414-1417.
9. Charbonnier J, van Rikxoort E, Setio A, Schaefer-Prokop C, van Ginneken B, Ciompi F. Improving airway segmentation in computed tomography using leak detection with convolutional networks. Med Image Anal. 2017;36:52–60.
10. Grinsven MJ, Hoyng CB, Theelen T, Sanchez CI. Fast convolutional neural network training using selective data sampling: Application to hemorrhage detection in color fundus images. In: Proc. IEEE Trans Med Imaging. 2016;35:1273-1284.
11. Karimi D, Samei G, Kesch C, Nir G, Salcudean SE. Prostate segmentation in mri using a convolutional neural network architecture and training strategy based on statistical shape models. Int J Comput Assist Radiol Surg. 2018;13(8):1211–1219.
12. Tran GS, Nghiem TP, Nguyen VT, Luong CM, Burie J-C. Improving accuracy of lung nodule classification using deep learning with focal loss. Int'l J Healthcare Eng. 2019;1-9.
13. Bejnordi BE, Veta M, van Diest P.J. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 2017;318(22):2199–2210.
14. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Proc Nat. 2017;542(7639):115-118
15. Yang W, Chen Y, Liu Y, Zhong L, Qin G, Lu Z, Feng Q, Chen W. Cascade of multi-scale convolutional neural networks for bone suppression of chest radiographs in gradient domain. Med Image Anal. 2016;35:421–433.
16. Litjens G, Kooi T, Bejnordi BE, Arindra A, Setio A, Ciompi F, Ghafoorian M, van der Laak JA, van Ginneken B, Sanchez CI. A survey on deep learning in medical image analysis. In: Diagnostic Image Analysis Group. 2017;1-38.
17. Hu K, Zhang Z, Niu X, Zhang Y, Cao C, Xiao F, Gao X. Retinal vessel segmentation of color fundus images using multiscale convolutional neural network with an improved cross-entropy loss function. J Neurocomput. 2018;309:179–191.
18. Chen X, Xu Y, Wong DWK, Wong TY, Liu J. Glaucoma detection based on deep convolutional neural network. Med Biol Soc: Proc. IEEE Int'l Conf; 2015;715-718.
19. Havaei M, Davy A, Warde-Farley D. et al. Brain tumor segmentation with deep neural networks. Med Image Anal. 2017;35:18–31.
20. Huynh BQ, Li H, Giger ML. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. Med Imaging 2016;3(3):034501.
21. Kamnitsas K, Ledig C, Newcombe VF, Simpson JP, Kane AD, Menon DK, Rueckert D, Glocker B. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. Proc Med Image Anal. 2017;36:61-78.
22. Kawahara J, BenTaieb A, Hamarneh G. Deep features to classify skin lesions. IEEE Int Symp Biomed Imaging. 2016;1397-1400.
23. Akram SU, Kannala J, Eklund L, Heikkila J. Cell segmentation proposal network for microscopy image analysis. In: Second International Workshop, DLMIA: Proc; 2016;21-29.
24. Cohen A, Rivlin E, Shimshoni I, Sabo E. Memory based active contour algorithm using pixel-level classified images for colon crypt segmentation. Med Imaging Graph. 2015;43:150–164.
25. Thoma M. A survey of semantic segmentation. 2016;1-16. arXiv:abs/1602.06541.

26. Guo Y, Liu Y, Georgiou T, Lew MS. A review of semantic segmentation using deep neural networks. Int'l Journal of Multimedia Information Retrieval. 2018;787-790.
27. Liu X, Deng Z, Yang Y. Recent progress in semantic image segmentation. Proc Artif Intell Rev 2018;52(2):1089–1106.
28. Goceri E. Challenges and recent solutions for image segmentation in the era of deep learning. Tools and Applications (IPTA): Proc.in Ninth Int'l conference on Image Processing Theory. 2019;1-6.
29. Taghanaki SA, Abhishek K, Cohen JP, Cohen-Adad J, Hamarneh G. "Deep semantic segmentation of natural and medical images: a review. CoRR. 2019; 54(1):137-178.
30. Siddique I, Bajwa I, Naveed M, Choudhary M. Auto-matic functional brain mr image segmentation using region growing and seed pixel. IEEE Int'l Conf. on Information and Communications Technology. 2006;1-12.
31. Zhu SC, Guo YWCE, Xu Z. What are textons? . Int'l Journal of Comput Vision. 2005;62:121-143.
32. Ho TK. Random decision forests. Document Analysis and Recognition: Proc. IEEE Int'l Conf; 1995;278-282.
33. Plath N, Toussaint M, Nakajima S. Multiclass image segmentation using conditional random fields and global classification. In: Proceedings of the 26th Annual International Conference on Machine Learning. (ACM). 2009;817-824.
34. Cheng H, Jiang X, Sun Y, Wang J. Color image segmentation: advances and prospects. Pattern Recogn. 2001;34(12):2259–2281.
35. Dalal N, Triggs B. Histograms of oriented gradients for human detection. Comput Vis Pattern Recogn 2005;886-893.
36. Lowe D. Distinctive image features from scale invariant keypoints. Int J Comput Vis. 2004;60:91-110.
37. Pietikäinen M, Mäenpää T, Viertola J. Color texture classification with color histograms and local binary patterns. Workshop on Texture Analysis in Machine Vision. 2002;1-4.
38. Bradski G, Pisarevsky V. Intel's computer vision library: applications in calibration, stereo segmentation, tracking, gesture, face and object recognition. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR, 2000.
39. Chen H, Zheng Y, Park JH, Heng PA, Zhou SK. Iterative multidomain regularized deep learning for anatomical structure detection and segmentation from ultrasound images. Med Image Comput Assist Interv: Proc. 2016;9901:487-495.
40. Brox T, Bourdev L, Maji S, Malik J. Object segmentation by alignment of poselet activations to image contours. Computer Vision and Pattern Recognition: IEEE Int'l Conf; 2011;2225-2232.
41. Farag A, Lu L, Roth HR, Liu J, Turkbey E, Summers RM. A bottom-up approach for pancreas segmentation using cascaded superpixels and (deep) image patch labeling. 2015;1-14. arXiv:1505.06236
42. Saidin N, Ngah UK, Sakim HAM, Siong DN, Hoe MK. Density based breast segmentation for sammograms using graph gut techniques. In TENCON 2009, 2009.
43. Adam A, Ioannidis C. Automatic road-sign detection and classification based on support vector machines and hog descriptors. Remote Sensing and Spatial Information Sciences: ISPRS Annals of the Photogrammetry. 2014;1-7.
44. Yang MY, Forstner W. A hierarchical conditional random field model for labeling and classifying images of man-made scenes. In 2011:196-203.
45. Korc F, Forstner W. etrims image database for interpreting images of man-made scenes. In: TR-IGG-P-2009-01. Department of Photogrammetry: University of Bonn; 2009.
46. Shotton J, Winn J, Rother C, Criminisi A. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. Computer Vision-ECCV: Springer; 2006;1-15.
47. Vemulapalli R, Tuzel O, Liu M-Y, Chellappa R. Gaussian conditional random field network for semantic segmentation. Computer Vision and Pattern Recognition: IEEE Int'l Conf; 2015;3224-3233.
48. Gulsrud TO, Engan K, Hanstveit T. Watershed segmentation of detected masses in digital mammograms. In Proceedings of the IEEE Conference on Engineering in Medicine and Biology 27th Annual Conference. 2005;3305-3307.
49. Huang YL, Chen DR. Watershed segmentation for breast tumor in 2D sonography. Ultrasound Med Bio. 2004;30:625-632.
50. Gomez W, Leija L, Pereira WCA, Infantosi AFC. Segmentation of breast, nodules on ultrasonographic images based on marke d-controlled watershed transform. Computación y Sistemas: Proc; 2010;14:165-174.
51. Pan Z, Lu J. A bayes-based region-growing algorithm for medical image segmentation. Comput Sci Eng. 2007;9(4):32–38.
52. Machine learning: An algorithmic perspective. CRC Press, 2015.
53. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille A. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. 2016;834-848. arXiv:1606.00915.
54. Kim H, Hwang S. Scale-invariant feature learning using deconvolutional neural networks for weakly-supervised semantic segmentation. 2016;1-17. arXiv:1602.04984
55. Krizhevsky A, Sutskever I, Hinton G. Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst. 2012;1097-1105.
56. Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol Cybernet. 1980;36:193-202.
57. Deep learning. MIT Press, 2016.
58. Lo S-CB, Chan H-P, Lin J-S, Li H, Freedman MT, Mun SK. "Artificial convolution neural network for medical image pattern recognition. In: Proceedings Neural Networks, 1995.
59. Yann L, Cortes C, Burges CJ. Mnist handwritten digit database. 2013.
60. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. Computer Vision and Pattern Recognition: Proc. IEEE Int'l Conf; 2015;79(10) 1337-1342.
61. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. 2017;1-8.
62. Jia Y. Caffe: An open source convolutional architecture for fast feature embedding. 2013. https://caffe.berkeleyvision.org.
63. Silberman N, Hoiem D, Kohli P, Fergus R. Indoor segmentation and support inference from rgbd images. In: European Conference on Computer Vision, Springer, 2012;746-760.
64. Liu C, Yuen J, Torralba A. Nonparametric scene parsing: label transfer via dense scene alignment. Computer Vision and Pattern Recognition: Proc. IEEE Int'l Conf; 2009;1972-1979.
65. Shelhamer E, Long J, Darrell T. Fully convolutional models for semantic segmentation. In: Pattern Analysis and Machine Intelligence: IEEE Trans; 2016;1-12.
66. Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In ECCV, 2018.
67. Bengio Y. Learning deep architectures for ai. In: Foundations and trends in machine learning, 2009.
68. Pinheiro PH, Collobert R. Recurrent convolutional neural networks for scene parsing. 2013;1-14. arXiv:1306.2795Ronne

69. Gould S, Fulton R, Koller D. Decomposing a scene into geometric and semantically consistent regions. In: IEEE 12th International Conference on Computer Vision, 2009: 1-8.
70. Ren X, Malik J. Learning a classification model for segmentation. In: Proceedings of the Ninth IEEE International Conference on Computer Vision, 2003;2:1-10.
71. Farabet C, Couprie C, Najman L, LeCun Y. Learning hierarchical features for scene labeling. In: Pattern Analysis and Machine Intelligence: IEEE Trans; 2013;1-15.
72. Sharma A, Tuzel O, Liu MY. Recursive context propagation network for semantic segmentation. NIPS, 2014.
73. Hong S, Noh H, Han B. Decoupled deep neural network for semi-supervised semantic segmentation. 2015. arXiv:1506.04924.
74. Lempitsky V, Vedaldi A, Zisserman A. A pylon model for semantic segmentation. In: Advances in Neural Information Processing Systems. 2011.
75. Kallenberg M, Petersen K, Nielsen M, Ng A, Diao P, Igel C, Vachon C, Holland K, Karssemeijer N, Lillholm M. "Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring," Proc. IEEE Trans Med Imaging 2016;35(5):1322-1331.
76. Zhu X, Goldberg A. Introduction to semisupervised learning. In: Synthesis lectures on artificial intelligence and machine learning 2009;3.
77. Roth H, Oda M, Shimizu N, Oda H, Hayashi Y, Kitasaka T, Fujiwara M, Misawa K, Mori K. "Towards dense volumetric pancreas segmentation in ct using 3d fully convolutional networks," Medical Imaging. 2017;1-6. arXiv:1711.06439
78. Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzago PA. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. J Mach Learn Res. 2010.
79. Janowczyk A, Basavanhally A, Madabhushi A. Stain normalization using sparse autoencoders (stanosa): Application to digital pathology. In: Proc: Comput Med Imaging Graph, In press; 2016: 3320-3328.
80. Chen M, Shi X, Zhang Y, Wu D, Guizani M. Deep feature learning for medical image analysis with convolutional autoencoder neural network. IEEE Trans Big Data. 2016;1-10.
81. Gondara L. Medical image denoising using convolutional denoising autoencoders. In: Proc. IEEE Int'l Conf. on Data Mining Workshops. 2016;242-246.
82. Hinton G. A practical guide to training restricted boltzmann machines. In: UTML TR 2010–003. Department of Computer Science: University of Toronto; 2010.
83. Brosch T, Traboulsee A, Li DK, Tam R. Deep learning of image features from unlabeled data for multiple sclerosis lesion segmentation. In: International Workshop on Machine Learning in Medical Imaging, Springer. 2014;117-124.
84. Pereira S, Meier R, McKinley R, Wiest R, Alves V, Silva CA, Reyes M. Enhancing interpretability of automatically extracted machine learning features: application to a rbm-random forest system on brain lesion segmentation. Med Image Anal. 2018;44:228–244.
85. Nahid A-A, Mikaelian A, Kong Y. Histopathological breast-image classification with restricted boltzmann machine along with backpropagation. Biom Res. 2018;29(10):2068–2077.
86. Cao P, Liu X, Bao H, Yang J, Zhao D. Restricted boltzmann machines based oversampling and semi-supervised learning for false positive reduction in breast cad. Bio-Med Mater Eng. 2015;26(s1):S1541–S1547.
87. G. E. Hinton. Deep belief networks. Scholarpedia. 2009;4(5):59472009.
88. Oquab M, Bottou L, Laptev I, Sivic J. Learning and transferring mid-level image representations using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014;1717-1724.
89. Shie CK, Chuang C-H, Chou C-N, Wu M-H, Chang EY. Transfer representation learning for medical image analysis. In: 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2015;711-714.
90. Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks ?. In: Advances in neural information processing systems. 2014;3320-3328.
91. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. Computer Vision and Pattern Recognition: Proc. IEEE Int'l Conf; 2009;1-11.
92. Singh S, Ho-Shon K, Karimi S, Hamey L. Modality classification and concept detection in medical images using deep transfer learning. In: International Conference on Image and Vision Computing, (IVCNZ), 2018;1-6.
93. Luc P, Couprie C, Chintala S. Semantic segmentation using adversarial networks. In: Workshop on Adversarial Training, NIPS 2016;1-9.
94. Li Y, Qi H, Dai J, Ji X, Wei Y. Fully convolutional instance-aware semantic segmentation. 2016;2359-2367. arXiv:abs/1611.07709.
95. Dai J, He K, Sun J. Instance-aware semantic segmentation via multi-task network cascades. 2015;3150-3158. arXiv:abs/1512.04412.
96. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014;1-14. arXiv:1409.1556.
97. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015;1-9.
98. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016;770-778.
99. Visin F, Kastner K, Cho K, Matteucci M, Courville AC, Bengio Y. Renet: a recurrent neural network based alternative to convolutional networks. 2015: 1-9.
100. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. Computer Vision and Pattern Recognition: Proc. IEEE Int'l Conf; 2015.
101. Badrinarayanan V, Kendall A, Cipolla R. Segnet: a deep convolutional encoder-decoder architecture for image segmentation. 2016;1-14. arXiv:1511.00561v3
102. Zhao H, Shi J, Qi X, Wang X, Jia J. "Pyramid scene parsing network. 2016;2881-2890. arXiv:abs/1612.01105
103. Kalesnykiene V, Kamarainen Jk, Voutilainen R, Pietilä J, Kälviäinen H, Uusitalo H. Diaretdb1 diabetic retinopathy database and evaluation protocol. 2014;1-10.
104. Porwal P, Pachade S, Kamble R, Kokare M, Deshmukh G, Sahasrabuddhe V, Meriaudeau F. Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. MDPI Data 2018;3(3):25.
105. Setio AAA, Jacobs C, Gelderblom J, van Ginneken B. Automatic detection of large pulmonary solid nodules in thoracic CT images. Med Phys. 2015; 42(10):5642–5653.
106. Cui S, Mao L, Jiang J, Liu C, Xiong, S. Automatic semantic segmentation of brain gliomas from MRI images using a deep cascaded neural network. Hindawi J Healthcare Eng. 2018;1-14.
107. Hein LM, Mersmann S, Kondermann D, Bodenstedt S, Sanchez A, Stock C, Kenngott HG, Eisenmann M, Speidel S. Can masses of non-experts train highly accurate image classifiers? In: Proc. Medical Image Computing and Computer-Assisted Intervention-MICCAI: Springer; 2014;438-445.
108. Coelho LP, Shariff A, Murphy RF. Nuclear segmentation in microscope cell images: a hand segmented dataset and comparison of algorithms. In: Proc. IEEE Int'l Symposium on Biomedical Imaging From Nano to Macro. 2009;518-521.

109. Sirinukunwattana K, Pluim JPW, Chen H, Qi X, Heng P, Guo YB, Wang LY, Matuszewski BJ, Bruni E, Sanchez U, Böhm A, Ronneberger O, Cheikh BB, Racoceanu D, Kainz P, Pfeiffer M, Urschler M, Snead DRJ, Rajpoot NM. Gland segmentation in colon histology images: the glas challenge contest. 2016. arXiv:1603.00275
110. Maska M, Ulman V, Svoboda D, Matula P. A benchmark for comparison of cell tracking algorithms. Proc Bioinform. 2014;30(11):1609–1617.
111. Arteta C, Lempitsky V, Noble J, Zisserman A. Learning to detect cells using non-overlapping extremal regions. In: MICCAI 2012, Part I. LNCS, 2012;348-356.
112. Kainz P, Urschler M, Schulter S, Wohlhart P. You should use regression to detect cells. In: MICCAI 2015. 2015;9351:276-283.
113. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers R. Chest x-ray: Hospital-scale chest x-ray database and benchmarks on weakly supervised classification and localization of common thorax diseases. In: Computer Vision and Pattern Recognition: IEEE Int'l Conf. 2017;3462-3471.
114. Aew J, Pollard T, Berkowitz S, Greenbaum N, Lungreen M, Deng C, Mark R, Horng S. Mimic-csr : a large database of labeled chest radiographs. 2019;1-7.
115. Shiraishi J, Katsuragawa S, Matsumoto T, Kobayashi T, Ichi Komatsu K, Matsui M, Fujita H, Kodera Y, Doi K. Development of a digital image database for chest radiographs with and without a lung nodule. Am J Roentgenol. 2000;174(1):71–74.
116. van Ginneken B, Stegmann M, Loog M. Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. Med Image Anal. 2006;10(1):19–40.
117. Chilamkurthy S, Ghosh R, Tanamala S, Biviji M, Campeau N, Venugopal V, Mahajan V, Rao P, Warier P. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. In: Proc. The Lancet. 2018;932:2388-2396.
118. Grammatikopoulou M, Flouty E, Kadkhodamohammadi A, Quellec G, Chow A, Nehme J, Luengo I, Stoyanov D. Cadis: cataract dataset for image segmentation. 2019;1-8.
119. Ye M, Giannarou S, Meining A, Yang G-Z. Online tracking and retargeting with applications to optical biopsy in gastrointestinal endoscopic examinations. Med Image Anal. 2015;30:144–157.
120. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M. et al., Tensorflow: Large-scale machine learning on heterogeneous distributed systems. 2016. arXiv:1603.04467
121. Bastien F, Lamblin P, Pascanu R, Bergstra J, Goodfellow I, Bergeron A, Bouchard N, Warde-Farley D, Bengio Y. Theano: new features and speed improvements. Int J Mach Learn. 2012:1-10.
122. Collobert R, Weston J, Karlen M. Natural language processing (almost) from scratch. 2011;12:2493-2537.
123. Chollet F. Keras. 2015. https://github.com/fchollet/keras.
124. Chen T, Li M, Li Y, Lin M, Wang N, Wang M, Xiao T, Xu B, Zhang C, Zhang Z. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. 1995;1-6.
125. Seide F, Agarwal A. Cntk: Microsoft's pen-source deep-learning toolkit. 2016.
126. Sharma A, Tuzel O, Jacobs DW. Deep hierarchical parsing for semantic segmentation. In: Computer Vision and Pattern Recognition: IEEE Int'l Conf. 2015;530-538.
127. Roth HR, Shen C, Oda H, Oda M, Hayashi Y, Misawa K, Mori K. Deep learning and its application to medical image segmentation. Med Imaging. 2018;1-6. arXiv:1803.08691v1
128. Smistad E, Lovstakken L. Vessel detection in ultrasound images using deep convolutional neural networks. In: Proceedings DLMIA. Vol. 10008 of Lect Notes Comput Sci. 2016;30-38.
129. Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, Liang J. Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE Trans Med Imaging. 2016;35(5):1299-q312
130. Zhoua X-Y, Shena M, Rigab C, Yanga G-Z, Lee S-L. Focal FCN: towards small object segmentation with limited training data. 2017. arXiv:1711.01506.
131. Lin, T-Y, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. In: Proc. IEEE International Conference on Computer Vision. 2017;2980-2988.
132. Zhoua X, Takayama R, Wang S, Hara T, Fujita H. Deep learning of the sectional appearances of 3d ct images for anatomical structure segmentation based on an FCN voting method. In: Med Phys 2017;44(10):5221–5233
133. Cicek O, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. 2016;424-432.
134. Ahn BB. The compact 3d convolutional neural network for medical images. Standford University, Stanford. 2017.
135. Hammack D. Forecasting lung cancer diagnoses with deep learning. In: Data Science Bowl 2017 Technical Report. 2017;1-6.
136. Jahangard S, Zangooei MH, Shahedib M. U-Net based architecture for an improved multiresolution segmentation in medical images. Electric Eng Syst Sci. 2020;1-22. arXiv:2007.08238
137. Lou A, Guan S, Loew M. DC-UNet: rethinking the u-net architecture with dual channel efficient CNN for medical images segmentation. Electric Eng Syst Sci. 2020;1-16. arXiv:2006.00414
138. Lei Y, Liu Y, Dong X, Tian S, Wang T, Jiang X, Higgins K, Beitler JJ, Yu DS, Liu T, Curran WJ, Fang Y, Yang X. Automatic multi-organ segmentation in thorax CT images using u-net-gan. In: Proc.SPIE Medical Imaging. 2019;10950.
139. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans Med Imaging. 2020;39(6):1856-1867.
140. Li X, Wang Y, Tang Q, Fan Z, Yu J. Dual unet for the segmentation of overlapping glioma nuclei. IEEE Access. 2019;7:84040–84052.
141. Yu L, Chen H, Dou Q, Qin J, Heng P-A. Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE Trans Med Imaging. 2016;36(4):994–1004.
142. Chen H, Dou Q, Yu L, Heng P-A. Voxresnet: Deep voxelwise residual networks for volumetric brain segmentation. NeuroImage. 2018;170:446-455.
143. Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using RNN encoder–decoder for statistical machine translation. in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014;1724-1734.
144. J. Chen, L. Yang, Y. Zhang, M. Alber, and D. Z. Chen, "Combining fully convolutional and recurrent neural networks for 3D biomedical image segmentation," 29th Conference on Neural Information Processing Systems (NIPS 2016), 2016;1-9.
145. Stollenga MF, Byeon W, Liwicki M, Schmidhuber J. Parallel multi-dimensional lstm, with application to fast biomedical volumetric image segmentation. 2015;1-13. arXiv:abs/1506.07452.
146. Poudel RPK, Lamata P, Montana G. Recurrent fully convolutional neural networks for multi-slice MRI cardiac segmentation. 2016;1-12.
147. Radau P, Lu Y, Connelly K, Paul G, Dick A, Wright G. Evaluation framework for algorithms segmenting short axis cardiac MRI. 2009.

148. Feng X, Yang J, Laine AF, Angelini ED. Discriminative localization in CNNs for weakly-supervised segmentation of pulmonary nodules. 2017;1-8. arXiv:abs/1707.01086
149. Gu Z, Cheng J, Fu H, Zhou K, Hao H, Zhao Y, Zhang T, Gao S, Liu J. Ce-net: Context encoder network for 2d medical image segmentation. IEEE Trans Med Imaging. 2019;38(10):2281–2292.
150. Oktay O, Ferrante E, Kamnitsas K, Heinrich M, Bai W, Caballero J. Anatomically constrained neural networks (ACNN): application to cardiac image enhancement and segmentation. IEEE Trans Med Imaging. 2018;37(2):384–395.
151. Alex V, Vaidhya K, Thirunavukkarasu S, Kesavadas C, Krishnamurthia G. Semisupervised learning using denoising autoencoders for brain lesion detection and segmentation. J Med Imaging. 2017;4(4):041311.
152. Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks?. 2014;1-9. arXiv:abs/1411.1792
153. Ravishankar H, Sudhakar P, Venkataramani R, Thiruvenkadam S, Annangi P, Babu N, Vaidya V. Understanding the mechanisms of deep transfer learning for medical images. 2017;1-8. arXiv:abs/1704.06040
154. Chen S, Ma K, Zheng Y. Med3d: Transfer learning for 3D medical image analysis. 2019;1-12. arXiv:abs/1904.00625.
155. Xue Y, Xu T, Zhang H, Long LR, Huang X. Segan: adversarial network with multi-scale loss for medical image segmentation. 2017;1-9. arXiv:abs/1706.01805.
156. Rezaei M, Yang H, Meinel C. Recurrent generative adversarial network for learning imbalanced medical image semantic segmentation. Multimed Tools Appl. 2019;79(21):15329–15348.
157. Jiang F, Grigorev A, Rho S, Tian Z, Fu Y, Jifara W, Adil K, Liu S. Medical image semantic segmentation based on deep learning, In: Neural Computing in Next Generation Virtual Reality Technology. 2017;1257-1265.
158. Cai J, Lu L, Zhang Z, Xing F, Yang L, Yin Q. Pancreas segmentation in MRI using graph-based decision fusion on convolutional neural networks. Med Image Comput Assist Interv. 2016;9901:442-450.
159. Thong W, Kadoury S, Piche N, Pal CJ. Convolutional networks for kidney segmentation in contrast-enhanced CT scans. In: Proceedings Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization. 2016;1-6.
160. Lessmann N, Isgum I, Setio AA, de Vos BD, Ciompi F, de Jong PA, Oudkerk M, Viergever Mali WPTMMA, Ginneken, B. Deep convolutional neural networks for automatic coronary calcium scoring in a screening study with low dose chest CT. In: Proc. Medical Imaging. Vol. 9785 of Proceedings of the SPIE, 2016. 1-6.
161. Juan J, Gomez Valverde GF, Anton Alfonso. Automatic glaucoma classification using color fundus images based on convolutional neural networks and transfer learning. In: Proceedings Biomedical Optics Express. 2019;10(2):892-913.
162. Li Z, MD Y, He S, Keel W, Chang Meng RT, He M. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. In: Proceedings American Academy of Opthulmology. 2018;125(8):1199-1206.
163. Raghavendra U, Fujita H, Bhandary SV, Gudigar A, Tan JH, Acharya UR. Deep convolution neural network for accurate diagnosis of glaucoma using digital fundus images. In: Proc. International Journal of Informatics and Computer Science Intelligent Systems Applications. 2018;441:41-49.
164. Dong FLYMH, Yang G, Guo Y. Automatic brain tumor detection and segmentation using u-net based fully convolutional networks. Medical Image Understanding and Analysis, MIUA: Proc; 2017;1-12.
165. Dvorak P, Menze B. Structured prediction with convolutional neural networks for multimodal brain tumor segmentation. MICCAI-BRATS: Proc; 2015; 13-24.
166. Alansary A, Kamnitsas K, Davidson A, Khlebnikov R, Rajchl M, Malamateniou C, Rutherford M, Hajnal JV, Glocker B, Rueckert D, Kainz B. Fast fully automatic segmentation of the human placenta from motion corrupted MIR. In: Med Image Computation Assist Interv: Proc; 2016;9901:589-597.
167. Gao Y, Maraci MA, Noble JA. Describing ultrasound video content using deep convolutional neural networks. In: IEEE Int Symp Biomedical Imaging: Proc; 2016;787-790.
168. Li Y, Ping W. Cancer metastasis detection with neural conditional random field. In: 1st Conference on Medical Imaging with Deep Learning (MIDL). 2018;1-9.
169. Kraus OZ, Ba JL, Frey BJ. Classifying and segmenting microscopy images with deep multiple instance learning. In: Bioinformatics. 2016;32(12): 152-159.
170. Birenbaum A, Greenspan H. Longitudinal multiple sclerosis lesion segmentation using multi-view convolutional neural networks. Second International Workshop, DLMIA: Proc; 2016;58-67.
171. Fotin SV, Yin Y, Haldankar H, Hofmeister JW, Periaswamy S. Detection of soft tissue densities from digital breast tomosynthesis: comparison of conventional and deep learning approaches. Medical Imaging(SPIE): Proc. 2016;9785:1-6.
172. Ramaswamy S, Truong K. Pulmonary nodule classification with convolutional neural networks. 2016. https://cs231n.stanford.edu/reports/2016/pdfs/324_Report.pdf
173. Avendi MR, Kheradvar A, Jafarkhani H. Automatic segmentation of the right ventricle from cardiac MRI using a learning-based approach. Magn Reson Med. 2016;78(6):2439–2448.
174. Guo Y, Wu G, Commander LA, Szary S, Jewells V, Lin W, Shent D. Segmenting hippocampus from infant brains by sparse patch matching with deep-learned features. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. 2014;7:87-93.
175. Mansoor A, Cerrolaza J, Idrees R, Biggs E, Alsharid M, Avery R, Linguraru MG. Deep learning guided partitioned shape model for anterior visual path- way segmentation. Imaging: IEEE Trans. Med; 2016;35(8):1856-1865.
176. Su H, Xing F, Kong X, Xie Y, Zhang S, Yang L. Robust cell detection and segmentation in histopathological images using sparse reconstruction and stacked denoising autoencoders. In: Lecture Notes in Computer Science, 9351. Springer; 2018;9351 383-390.
177. Cai Y, Landis M, Laidley DT, Kornecki A, Lum SLA. Multimodal vertebrae recognition using transformed deep convolution network. Comput Med Imaging Graph. 2016;51:11-19.
178. Azizi S, Imani F, Ghavidel S, Tahmasebi A, Kwak JT, Xu S, Turkbey B, Choyke P, Pinto P, Wood B, Mousavi P, Abolmaesumi P. Detection of prostate cancer using temporal sequences of ultrasound data: a large clinical feasibility study. Surgury 2016;11(6):947-956.
179. C. Shorten and T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. J Big Data (2019) 6:60, 2019.
180. Souly N, Spampinato C, Shah M. Semi supervised semantic segmentation using generative adversarial network. In 2017: 5688-5696.
181. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014; 15(1):1929–1958.
182. Hendrycks D, Gimpel K. Adjusting for dropout varience in batch normalization and weight initialization. 2016;1-10.
183. Erhan D, Bengio Y, Courville A, Manzagol P-A, Vincent P. Why does unsupervised pre-training help deep learning? J Mach Learn Res 2010;11:201–208.

184. Sulimov P, Sukmanova E, Chereshnev R, Kertesz-Farkas Guided layer-wise learning for deep models using side information 2019;191102048:1-12.

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