






Medical Image Semantic Segmentation Using Deep Learning: A Survey

Ferialle Lahreche¹ , Abdelouahab
Moussaoui² , and Slimane
Oulad-Naoui^{1,*} 

¹ Lab. des Mathématiques et Sciences Appliquées, University of Ghardaia, Algeria.

² Department of Computer Science, Ferhat Abbas University, Sétif, Algeria.

* Corresponding author: SLimane Oulad-Naouis.ouladnaoui@gmail.com

Abstract. Biomedical image segmentation has witnessed a significant advancement with the emergence of deep learning (DL) technologies, which become pivotal in medical image analysis. This paper presents a comprehensive review of the evolution and current state of medical image segmentation (MIS) techniques, with a particular focus on semantic segmentation enabled by DL. We conduct a critical analysis of various neural network architectures, including the latest developments in vision transformers, and their impact on enhancing the accuracy and efficiency of medical image processing. We identify key advancements, discuss current challenges, and suggest potential future directions.

Keywords: Medical Imaging, Semantic Segmentation, Deep Learning

1 Introduction

Image processing, analysis, and understanding techniques are an integral part of numerous applications and an active research area in computer vision. Image segmentation, in particular, is a highly relevant and widely researched topic, with applications in numerous domains, including the medical field.

Semantic segmentation of medical images and the specific localization of lesions from medical images of various types is a very important area of focus, given the sensitivity and importance of this area. In the past, medical images were segmented to determine the shapes and sizes of organs, as well as the types and locations of tumors within them, based on traditional methods that use domain knowledge to achieve good segmentation and rely heavily on feature selection such as Thresholding (P-tile Method, Otsu Method), Edge-Based (Gradient Based Method and Gray Histogram Technique), and Region-Based techniques (Region Growing and Region Splitting and Merging). After a while, machine learning methods became the dominant technology for a long time such as Neural Networks, K-means, Support Vector Machine (SVM), but the need to accurately extract features manually was an obstacle to the development of these technologies [2, 13]. With the improvement of equipment and the advancement of medical treatment, all kinds of new medical imaging devices have become

© The Author(s) 2024

C. A. Kerrache et al. (eds.), *Proceedings of the International Conference on Emerging Intelligent Systems for Sustainable Development (ICEIS 2024)*, Advances in Intelligent Systems Research 184,

https://doi.org/10.2991/978-94-6463-496-9_25

more common, and the types of medical imaging widely used have evolved, the most prominent of which are computed tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET).), X-ray, and ultrasound (UI). With the abundance of very useful information in modern medical images, they have become the main basis for clinical diagnosis for doctors, but machine learning techniques can no longer keep up with them, which led to the emergence of DL techniques in the 2000s, which showed tremendous ability and surpassed all its predecessors in terms of segmentation accuracy and speed, greatly reducing the workload of doctors [2, 13]. This prompted us to undertake a comprehensive review of DL technologies, starting with the oldest and ending with the newest, focusing on the shortcomings of the technologies and the reasons for the emergence of new ones, as well as their strengths presenting some research papers that used every mentioned technique, where transformers, in particular hybrid transformers, are currently the subject of extensive research and are witnessing a significant increase in the number of publications.

This paper commences with an introduction that outlines the historical development of MIS techniques, before elucidating the various types of medical images and the associated MIS. This is then followed by the presentation of a number of published medical datasets, accompanied by an analysis of the most commonly employed evaluation metrics in the context of DL techniques. Subsequently, an explanation of DL techniques used for MIS was presented, with a focus on the most prominent published works. Furthermore, we present an explanation of DL network training techniques, with a particular emphasis on the most prominent challenges in this field and suggestions for addressing them. Finally, the research is concluded with a conclusion.

2 Medical Imaging

Medical imaging is a non-invasive technology whose objective is to create visual images of the internal tissues of the human body by acquiring signals through the utilisation of the physical principles of sound, light, electromagnetic waves, and so forth. Several widely used medical imaging modalities produce different types of medical images, including ultrasound, digital radiography, computed tomography (CT), magnetic resonance imaging (MRI), optical coherent tomography (OCT), mammograms, and positron emission tomography (PET) [19, 21].

1. **X-ray Imaging:** is a significant repository of medical data for use by doctors and researchers. X-ray images are obtained through the use of radiation that is part of the electromagnetic spectrum and which passes through the body to produce informative images of the tissues and internal structure of the body. X-rays are a relatively inexpensive, convenient, and widely applicable imaging modality in the field of medical imaging [18, 22].
2. **Computed Tomography (CT) Scans:** is a computerized X-ray imaging technique that employs a narrow beam of radiation, which is then rapidly rotated around the body in order to capture detailed internal images. This

process relies on the high contrast between gas and tissue [23, 22]. CT produces detailed cross-sections of bones, soft tissues, and organs that can be formatted in multiple frames to generate three-dimensional images of a segment of the body [18].

3. **Optical Coherence Tomography (OCT) Images:** employs light waves to create cross-sectional images of the retina, which can be utilized to examine the distinct layers of the retina that facilitate mapping and measurement of their thickness. It plays a pivotal role in the diagnosis of retinal diseases [22].
4. **Fundus Images:** is a sophisticated two-dimensional imaging modality that captures the rear of the eye and is employed for the diagnosis of a multitude of medical conditions, including the detection and classification of hypertensive retinopathy [22, 18].
5. **Magnetic Resonance Imaging (MRI):** is capable of capturing highly detailed two-dimensional and three-dimensional anatomical images due to its powerful, non-invasive, and effective imaging technology, which does not utilize radiation. This makes it the optimal choice for image capture when frequent imaging is required in the treatment process. In terms of image quality, it is considered the best available option; however, it is more expensive than X-rays and CT scanning and requires a longer examination time [18, 23, 22].
6. **Ultrasound:** is a non-invasive medical procedure that uses sound waves without any radiation to create an image of the internal organs, tissues, and other structures within the body [22].
7. **Histopathology or Whole-Slide Imaging (WSI):** refers to the capture of microscopic tissue samples from a biopsy glass slide or surgical specimen. This involves the acquisition of small, high-resolution squares or strips of images, which are then montaged to create a complete high-resolution digital image of the histological section. This image can be stored as efficiently high-resolution digital files, which can be accessed, analyzed, and shared with scientists via the Internet using slide management techniques [22].

3 Medical Image Segmentation

The first development of image segmentation techniques dates back to 1965, when the Roberts operator or Roberts edge detector was introduced in [3], which is the first step towards decomposing the image into its basic components, after which many techniques and algorithms followed [5].

Image Segmentation is defined as the process that divides an image into its component parts, objects, or non-overlapping regions that have certain properties where the level of this segmentation varies according to the problem to be solved and stops when the elements of interest are isolated.

More formally, suppose the image is represented by R , the segmentation of R for a uniformity predicate P is the division of R into disjoint non-empty parts R_i , where $i = 1, 2, \dots, n$ so that [4, 5]:

1. $\bigcup_{i=1}^n R_i = R$;
2. for all i and $j, i \neq j$ there exists $R_i \cap R_j = \emptyset$;
3. for $i = 1, 2, \dots, n$, it must have $P(R_i) = TRUE$;
4. for all $i \neq j$, there exists $P(R_i \cup R_j) = FALSE$;

Where \emptyset represents an empty set and $P(R_i)$ is a uniformity predicate for all elements in set R_i .

Others thought that the following condition is important also:

5. For all $i = 1, 2, \dots, n, R_i$, is a connected component.

The first condition means that the sum of the segmented regions contains all pixels in an image. The second condition means that the different segmented regions do not overlap. The third condition means that there are some similar properties between the pixels in the same segmented regions. The fourth condition refers to the difference in some properties between the pixels belonging to different segmented regions, and finally, the fifth condition means that the pixels in the same region are connected to each other [5].

Image Segmentation can be separated into two categories: **Instance Segmentation**, which distinguishes between different instances of the same object in addition to classification and localization, and **Semantic Segmentation**, which is a classification process for each pixel, where there are two challenges in this task that must be dealt with simultaneously, namely classification and localization. For the classification, each object associated with a specific semantic concept must be distinguished correctly, and in the localization, the classification of pixels must be aligned with the appropriate coordinates in the output score map [1, 2].

The process of analyzing and processing 2D or 3D images for segmentation, extraction, 3D reconstruction, and 3D display of human organs, soft tissues, and sick bodies using computer image processing technology is known as accurate MIS [6]. It is a crucial stage in the process of computer-assisted diagnosis, image-guided surgery, and treatment planning. Furthermore, early diagnosis of certain diseases can help to avoid serious consequences, such as permanent disabilities like vision loss, so methods have varied to meet this need, ranging from CNNs to Vit-Based, which are among the best image segmentation techniques.

MIS is different from other image segmentation since the background of medical scans is typically scattered and we are not only searching for objects in the image but also for different organs or the segmentation of an organ. MIS techniques are typically divided into two categories: organ-specific and multi-organ. This distinction is based on the differing levels of context modeling required by each approach [20].

- **Organ-Specific Segmentation:** This approach considers a certain feature of the underlying organ in the design of architectural components or loss functions. The approach is divided into two categories based on the type of input: 2D and 3D [20].

1. Brain: The field of brain disease analysis encompasses the detection of a range of conditions, including brain tumors, strokes, traumatic brain

injuries, brain metastases, Alzheimer’s disease, epilepsy, and other diseases. This is typically accomplished through magnetic resonance imaging (MRI), where the images are divided into pre-treatment and post-treatment scans. Each patient is examined using instruments with varying magnetic field intensities and protocols. Four main types of MR images can be distinguished: The imaging techniques employed in this context include T1, T1c, T2, and FLAIR [6, 24].

2. Eye: The eye is the most sensitive organ in the human body, and the segmentation of the blood image in the retina is a challenging process. There are numerous unsolved segmentation problems, including the loss of small and weak blood vessels and over-segmentation [6].
 3. Chest: In the case of chest lesions, X-rays are frequently employed as a diagnostic tool, as they are the most commonly used imaging technique. Chest lesions can be broadly categorized into several distinct groups, including lung lesions such as pneumothorax and pneumonia, cardiac lesions such as ventricular hypertrophy, and bony lesions such as rib fractures [24].
 4. Abdomen: Abdominal imaging is frequently conducted using computed tomography (CT) and magnetic resonance imaging (MRI) techniques, which permit the visualization of a range of abdominal structures, including solid organs such as the liver, kidneys, spleen, and pancreas, as well as the lower abdominal organs [6, 24].
- **Multi-organ Segmentation:** It attempts to segment multiple organs simultaneously, which is challenging due to inter-class imbalance and the diverse sizes, shapes, variances, and contrasts among the organs [20].

4 Medical Images Segmentation: Datasets and Evaluation Metrics

4.1 Medical Datasets

To achieve optimal results, DL models require a substantial quantity of data. In the current era, a plethora of data types are readily available, except medical data, which is subject to privacy regulations and the necessity for expert processing. Consequently, alternative techniques have been employed to compensate for this deficit, including data augmentation and generation, transfer learning, and so forth. Medical images are classified into three distinct categories: two-dimensional (2D) images, two-and-a-half-dimensional (2.5D) images (RGB images), and three-dimensional (3D) images [2].

Some Known Datasets are present in the following table 1.

4.2 Evaluation Metrics

As usual, any DL model needs evaluation metrics. Accuracy and precision are some such metrics, but in medical segmentation algorithms, doctors’ hand-drawn

Table 1. Popular Medical Images Segmentation Datasets.

Dataset Name	Organ	Modality	Dimension
Pancreas-CT	abdominal	CT	3D
3D-IRCADb-01	liver	CT	3D
Kaggles Data Science Bowl (DSB) 2017	Chest	CT	2D
LUnG Nodule Analysis LUNG16	Lung	CT	2D
SIIM-ACR Pneumothorax Kaggle	Chest	X-ray	2D
Synapse multi-organ CT	Multi-organ	CT	2D
LiTS2017	Liver	CT	2D
BraTS 2020	Brain	MRI	3D

annotations are usually used as the gold standard (ground truth, GT). Other results of the segmentation algorithm are the prediction results (Rseg, SEG), which means that accuracy or precision alone is not sufficient to determine the quality of the segmentation, other methods have been invented [6]. The popular metrics employed are represented in terms of the following [2]:

- **True positive (TP)** represents that the actual data class and the predicted data class are true.
- **True negative (TN)** represents that both the actual data class and the predicted data class are false.
- **False positive (FP)** represents that the actual data class is false while the predicted data class is true.
- **False negative (FN)** represents that the actual data class is true while the predicted data class is false.

Some famous evaluation metrics that were mentioned in the literature are:

1. **Intersection over Union (IoU):** IoU or Jaccard index is the amount of intersecting area between the predicted image segment and the ground truth mask, divided by them [2].

$$IoU = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

2. **Dice index or coefficient** Is twice the amount of intersection area between the segment predicted and the ground truth divided by the total number of pixels in both [2].

$$dice = \frac{2|A \cap B|}{|A| + |B|} \quad (2)$$

3. **Precision:** Is the proportion of input data cases that are reported to be true and represented [2].

$$PR = \frac{TP}{TP + FP} \quad (3)$$

4. **Recall:** Is the percentage of the total relevant results correctly classified by the model [2].

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

5. **F1 Score:** Is the harmonic average of the precision and recall values [2].

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (5)$$

6. **Pixel Accuracy:** Is the percentage of correctly classified pixels in the input image by the model [2].

$$Pixel\ Accuracy = \frac{\text{no. of pixels properly classified}}{\text{total number of pixels}} \quad (6)$$

5 Deep Learning Medical Images Segmentation Techniques

The advent of DL techniques and their demonstrated superiority over previous approaches has led to the development of a multitude of deep neural network architectures. Below we will mention those used in image segmentation.

Among the most powerful DL networks, we find **convolutional neural networks (CNN)**, which combine DL with image processing technology, which made it the leader in the field of image analysis and processing and attend many achievements, such as extracting image features, classifying them, and recognizing patterns where the fully convolutional network (FCN) was the first successful DL network for image semantic segmentation. It was regarded as the pioneering work of utilizing convolutional neural networks (CNNs) for image segmentation. Subsequently, other noteworthy segmentation networks were identified, including U-Net, Mask R-CNN, RefineNet, and DeconvNet. These networks demonstrated a notable proficiency in processing fine edges [6].

5.1 CNN

CNNs are supervised DL models that consist of several linked layers, each with a specific function, passing information and sharing weights of feature mapping in different locations between them to get the final output [13]. One of the earliest CNN architectures is LeNet5, proposed by Lecun et al. [25] which was successfully applied to handwriting recognition. Subsequently, a deeper network, AlexNet, was proposed in 2012 by Krizhevsky et al. [26], this network was the most successful at image classification of the ImageNet dataset then Zeiler and Fergus [27] subsequently presented ZFNet, a fine-tuning of the AlexNet structure. For 2D input images, ResNet [28] and VGGNet [29] were employed. Additionally, GoogleNet was introduced by Szegedy et al. [30] with an Inception module, and subsequently, Szegedy et al. [31] introduced two new modules, Inception V2 and Inception V3. In their work, Szegedy et al. [32] proposed

Inception-ResNet-v1, Inception-ResNet-v2, and a pure inception variant, inception V4. Chollet et al. [33] proposed a module named Xception, which translates as "extreme inception". There are other networks, such as SqueezeNet [34], DenseNet [35], and others. The distinction between these networks lies in the number of convolutions and pooling layers, with crucial process blocks situated between them. In other instances, even the concept of convolution and pooling is transformed [6, 2, 14].

In the study by Zhang et al. [36], deep convolutional neural networks (CNNs) were employed to segment brain tissue in the isointense stage into white matter (WM), grey matter (GM), and cerebrospinal fluid (CSF) using multi-modality MR images. The performance of the proposed approach was then compared to that of commonly used segmentation methods. The results demonstrated that the model significantly outperformed previous methods for segmenting infant brain tissue. The overall Dice ratios achieved for CSF, GM, and WM were 0.835, 0.852, and 0.864, respectively.

In [37], the authors proposed a system comprising three 2D CNNs for the segmentation of tibial cartilage in MRI scans of low field knee joints. The system was tested on 114 unseen scans. The model demonstrated superior performance compared to the state-of-the-art methods, utilizing 3D multi-scale features. The triplanar CNN was tested on 114 unseen scans, with a mean Dice Similarity Coefficient (DSC) of 0.8249, an accuracy of 99.93%, a sensitivity of 81.92% and a specificity of 99.97%. The standard deviation was 4.26% for DSC, 1.86% for accuracy, 7.62% for sensitivity, and 1.74% for specificity.

In [38] a single convolutional neural network (CNN) was trained to segment six tissues in brain magnetic resonance imaging (MRI), pectoral muscle in breast MRI, and coronary arteries in heart computed tomography angiography (CTA). The combined training procedure resulted in segmentation performance equivalent to that of a CNN that was trained specifically for that task for each.

Despite the excellent performance of CNN segmentation models, there are still limitations. One such limitation is the inability of fully connected layers to manage different input sizes. Additionally, it is not possible to use CNN with a fully connected layer for the object segmentation task, as the number of objects of interest in the image segmentation task is not fixed. Consequently, the length of the output layer cannot be fixed [2].

5.2 FCN

To overcome the limitations of CNN, Long et al. [39] proposed a new architecture called Fully Convolutional Network (FCN), which was the first DL network to be applied to the semantic segmentation of images. This achieved impressive results. The FCN architecture was inspired by the VGG network architecture. The main difference between it and CNN is that the last fully connected layer of the CNN is converted into a fully convolutional layer in the FCN, enabling it to accept input images of any size. Other significant alterations include the incorporation of a deconvolution layer, which serves to upsample the feature map of the final convolution layer and restore it to the same dimensions as the input image.

This process generates a dense pixel-wise prediction while preserving the spatial information present in the original input image. The final image segmentation was completed by performing pixel-by-pixel classification on the upsampled feature map. The FCN varies according to the upsampling to FCN-32s, FCN-16s, and FCN-8s [6, 2].

[40] utilized FCN on the same modalities and dataset as [36], and the results demonstrated the superiority of FCN over CNN. This was evidenced by the mean Dice coefficient of 0.885, in comparison to 0.864, achieved by the multi-FCNs (mFCNs) architecture, which involved training one network per modality, then fusing multiple-modality features from the high layer of each network demonstrated superior performance to the FCN model, which combined three modalities as input feature maps for networks. The average Dice ratios for the three modalities were 0.855 for CSF, 0.873 for GM, and 0.887 for WM, derived from eight subjects.

The authors in [41] trained a multi-class 3D FCN on manually labeled CT scans of seven abdominal structures (artery, vein, liver, spleen, stomach, gallbladder, and pancreas) on 281 clinical CT images and validated it on 50 clinical CT images. The models were tested on a completely unseen dataset obtained at a different hospital, which included 150 CT scans with three anatomical labels (liver, spleen, and pancreas). The hierarchical approach yielded a mean Dice score of 82.2%, representing a significant improvement over the previous mean of 68.5%. This represents the highest reported average score on the aforementioned dataset.

Other researchers employed variations of FCN, such as the fully convolutional DenseNet (FC-DenseNet), as described in [42]. This approach was used to segment and detect pneumothorax on chest X-rays, with the researchers achieving a mean pixel-wise accuracy (MPA) of 0.93 ± 0.13 and a dice similarity coefficient (DSC) of 0.92 ± 0.14 . The overall accuracy was 93.45%, with an F1 score of 92.97%.

The conventional FCN model is subject to certain limitations, including a tendency to be slow for real-time inference and a lack of efficiency in the incorporation of global context information. Additionally, it may result in the generation of low-resolution predictions with blurring in the object boundaries due to the downsampling of the resolution of feature maps generated at the output, which occurs as a consequence of the propagation of information through alternative convolution and pooling layers [2].

5.3 U-Net

The U-Net is an encoder-decoder network proposed by Ronneberger et al. [43] that employs the concept of deconvolution introduced by [27] and is constructed upon the FCN's architectural framework. The encoding component comprises convolutional and pooling layers, while the decoding component comprises alternating upsampling, which increases the size of a feature map, and pooling layers. The features are concatenated with the feature retained at the corresponding level in the encoder after the upsampling, thus combining the location

information obtained from the downsampling path with the contextual information obtained from upsampling. This structure enables the model to identify both the detailed and general features of objects, thus enabling pixel-wise classification for segmentation. Consequently, U-Net has become the gold standard for MIS, inspiring a plethora of meaningful enhancements [2, 24, 13, 44].

In [45], the authors employed a network based on the U-Net architecture for the segmentation of lungs in 2D chest X-rays, using both the original JSRT dataset and the BSE-JSRT dataset. The results demonstrate the high efficiency of the U-Net structure, particularly when utilizing GPUs.

In [46], the authors proposed an extension of the U-Net architecture to handle 3D data for volumetric segmentation. This was achieved by replacing all 2D operations with their 3D counterparts. The network was tested on highly variable 3D structures, including the *Xenopus* kidney, and achieved satisfactory results for both use cases.

The limitations of the U-Net model can be briefly described as follows: the input image size is constrained to 572, and the learning process is slow in the middle layers of the model, which causes the network to ignore the layers with abstract features [2]. Several variants of the U-Net architecture have been proposed, the most notable of which are: U-Net ++ [47], Attention U-Net [48] and SD-UNet [49].

5.4 V-Net

Although the performance of 3D U-Net was satisfactory, it is still limited in that it cannot effectively extract deep layer image features due to the constraints of its computational resources, which permit only three down-sampling operations. A variant, designated V-Net, was proposed by [50] that incorporates compression and decompression networks, as well as residual connections, which facilitate accelerated network convergence and prevent gradient vanishing. This results in a deeper network architecture and enhanced performance [2, 44].

In [51], the researchers used V-Net with a larger, multiscale receptive field for automatic segmentation of abdominal anatomy on CT images for 8-organ. They compared it to current DL methods and found that it yielded significantly higher dice scores for all organs and lower mean absolute distances for most organs, including Dice scores. The results were 0.78 Dice Score versus 0.71, 0.74, and 0.74, for the pancreas 0.90 versus 0.85, 0.87, and 0.83 for the stomach, and for the esophagus 0.76 versus 0.68, 0.69, and 0.66.

5.5 RNNs

Recurrent Neural Networks are designed to deal with sequences, where they are used in handwriting, speech recognition, and other natural language processing applications due to the recurrent connections that enable the network to memorize patterns from recent inputs. In addition, when used in computer vision tasks such as semantic segmentation and scene segmentation it also achieved very satisfactory results due to its ability to learn long-term dependencies from

the sequenced data and its ability to retain memory along the sequence [13–15]. The use of RNNs for segmentation was employed to model the time dependence of medical image sequences [44].

In [52], the authors proposed a recurrent convolutional neural network (RCNN) based on U-Net as well as a recurrent convolutional neural network (RRCNN) based on U-Net models, named RU-Net and R2U-Net respectively. The proposed models were tested on three different segmentation tasks: blood vessel segmentation from retinal images, skin cancer lesion segmentation, and lung segmentation from 2D images. The results demonstrated that the RU-Net and R2U-Net models exhibited superior performance in segmentation tasks with the same number of network parameters compared to existing methods including U-Net models and residual U-Net (or ResU-Net) models, in the three datasets.

In [53], the authors proposed a framework called FCSLSTM for 4D image segmentation. This framework utilizes FCNs for the spatial model and LSTM for the temporal model. When applied to segment the BRIC clinical dataset, which contains longitudinal pediatric magnetic resonance imaging (MRI), the approach yielded promising results.

It is worth noting that other networks have been developed, including Contextual LSTM (CLSTM) [54], recurrent FCN (RFCN) [55] created through the GRU application on the FCN system, and Clockwork RNN (CW-RNN) [56].

5.6 R-CNN

R-CNNs were employed for object detection and segmentation. The algorithm generates a region proposal network for bounding boxes through a selective search process. These region proposals are subsequently warped to standard squares and forwarded to a convolutional neural network (CNN), which generates a feature vector map as the final output. The output layer comprises features extracted from the image, which are then fed to the classification algorithm to classify the objects within the region proposal network. Additionally, the algorithm predicts offset values to enhance the precision of the region proposal [2]. The primary limitation of the RCN model is its inability to be implemented in real time, which may result in the generation of inaccurate candidate region proposals. As a result, alternative approaches were proposed, including fast R-CNN [57], faster R-CNN [58], and mask R-CNN [59].

5.7 DeepLab

The Deeplab model employs a pre-trained convolutional neural network (CNN) model, namely Resnet-101/VGG-16. Deeplabv1 [60] is based on the VGG network, with additional atrous convolution and conditional random field (CRF) employed to enhance the accuracy of segmentation boundaries. In the Deeplab V2 model [61], the authors replaced the Resnet-101 network with a more flexible use of atrous convolution, which they had previously employed in the atrous spatial pyramid pooling (ASPP) module. In the case of Deeplab V3 [62], the authors continued to utilize the Resnet-101 network with the parallel atrous convolution

module and ASPP module, while also beginning to remove CRF. In Deeplab V3+ [63], the ASPP unit from Deeplab V3 is combined with an encoder-decoder structure, utilizing the Xception model for semantic segmentation tasks [6, 2]. In their study, Tang et al. [64] proposed a method based on Faster R-CNN and DeepLab, which they termed the "detection and segmentation laboratory" (DSL) method. The Faster R-CNN was used to detect the liver area, and the results were then passed to DeepLab as input for segmentation. The authors evaluated their work on two datasets: 3Dircadb and MICCAISliver07. In comparison to the state-of-the-art automatic methods, the approach demonstrated superior performance, with the lowest volumetric overlap error observed on the. In addition, there are other works that may be of interest, including [65], and [66].

5.8 Transformer-Based-Techniques

In recent years, an architecture has emerged that outperforms its predecessors in both natural language processing and computer vision, called **Transformers** [16]. Transformers also known as language transformers, are feed-forward modules in neural networks that compute global representations and dependencies, and they are based on self-attention mechanisms. It first appeared as a machine translation architecture, and because of its effectiveness in numerous natural language processing applications, it has replaced recurrent models as the standard option. Because of its widespread use, researchers began considering how to apply it to computer vision applications. This resulted in the creation of **Vision Transformer (ViT)** [17], the first pure transformer for computer vision tasks, which made up for CNN networks' inability to capture long-range dependencies like the extraction of contextual information and non-local association of objects [18–20]. Due to their impressive results, as they scale up more easily and are more robust in the face of corruption, as the weak inductive bias that enables them to outperform CNN networks, has led to their use in segmenting medical images semantically since these images are more complex than the rest and results of applying CNN on medical images were limited [19].

Self-Attention The attention mechanism was inspired by the way humans observe and learn. When we observe and learn, we unintentionally pay attention to a part of the information while ignoring the rest [18]. Self-attention, also known as intra-attention, is a mechanism that defines relationships between data, regardless of their position in the sequence, thereby capturing long-term dependencies. The attention function takes a query and a set of key-value pairs as input vectors and assigns them to an output by computing the weighted sum of the values. The value weights are calculated as the attention of the query with its corresponding key [16, 20, 18].

Multi-Head Self Attention (MHSA) Rather than focusing on a single instance of attention, it has been demonstrated that it is more efficient to apply multiple instances of self-attention in parallel to the same input in order to capture hierarchical features. h self-attentions are applied where the output is the concatenation and projection of all those self-attention blocks [16, 18].

The original ViT was applied straight to the images with very little alteration.

The sequence of linear embeddings of the patches created from the divided image was fed into ViT. Tokens and image patches were managed in the same manner [17].

Subsequently, we will present a selection of works that employed pure or hybrid transformers and compare them with other structures, which demonstrated remarkable outcomes.

- **Boundary-Aware Transformer (BAT):** Boundary-Aware Transformer (BAT) was proposed by [67] and was used to address the challenges of automatic skin lesion segmentation. This was achieved by integrating a new boundary-wise attention gate (BAG) into transformers, which enabled the whole network to model global long-range dependencies and capture more local details by fully benefiting from boundary-wise prior knowledge. The BAT algorithm was trained on the ISIC 2016 dataset and evaluated on the PH2 dataset. It was then validated on the ISIC 2018 dataset. In the case of ISIC 2016 and PH2, the Dice coefficient achieved by BAT was 0.921, with an Intersection over Union (IoU) of 0.858. In the case of ISIC 2018, the Dice coefficient achieved by BAT was 0.912, with an IoU of 0.843.
- **TransBridge:** A lightweight model, designated "TransBridge," was developed by [68]. This model combines a CNN encoder-decoder structure with a transformer architecture, which was evaluated on the EchoNet Dynamic dataset for a left ventricular segmentation task. The results demonstrated a reduction in the total number of parameters and an improvement in the dice coefficient, which reached 91.4%.
- **COTRNet:** In their work [69], the authors proposed a convolutional-transformer network (COTRNet), which comprises a decoder and a decoder structure. The encoder is composed of several convolution-transformer blocks, while the decoder contains several up-sampling layers with skip connections from the encoder [20]. COTRNet was employed for the end-to-end segmentation of kidneys, kidney tumors, and kidney cysts in the 2021 challenge for kidney and kidney tumor segmentation (kits21). This approach yielded an average dice coefficient of 61.6%, a surface dice coefficient of 49.1%, and a tumor dice coefficient of 50.52%. The resulting segmentation ranked 22nd in the kits21 challenge.
- **TransBTS:** In [70], the authors proposed a network called TransBTS, which employs an encoder-decoder architecture with Transformer technology integrated into a 3D convolutional neural network (CNN) for the segmentation of brain tumors in magnetic resonance imaging (MRI). The encoder employs a 3D CNN to extract 3D volumetric spatial feature maps, while the transformer receives the reformulated feature maps for global feature modeling. The experimental results on the 2019 and 2020 BraTS datasets demonstrate that the model achieves comparable or superior outcomes to those of previous state-of-the-art 3D approaches for the same task. The authors of [71] were inspired by the TransBTS architecture and proposed the Bi-Transformer U-Net (BiTr-UNet), which consists of an attention module to refine encoder and

decoder features and two ViT layers. This approach was found to perform relatively better in the BraTS 2021 segmentation challenge.

- **Pure Transformer-Based Model:** In [72], the authors proposed a model based on the transformer architecture. Unlike traditional convolutional neural networks, this model does not involve any convolutional processes. Instead, it employs self-attention between adjacent image patches to predict the segmentation map of the central patch. The proposed model demonstrated superior segmentation accuracy compared to state-of-the-art convolutional neural networks (CNNs) in three data sets: The data sets comprised the following: Brain Cortical Panel T2 MRI, Pancreas CT, and Hippocampus MRI.

We end this section with a table summarizing all the previously mentioned networks, along with their most important characteristics and shortcomings:

Table 2. Medical Image Segmentation Approaches.

Approach	Architecture	Applications	Advantages	Limitations
CNN	CNN	Classification, Segmentation	Automatic feature extraction	Inability of fully connected layers to manage different input sizes
FCN	CNN	Segmentation	Generates a dense pixel-wise prediction while preserving the spatial information	Slow for real-time inference, Generate low-resolution predictions
U-Net	2D and 3D CNN	Segmentation	Identify both the detailed and general features of objects, Gives a height precision	Slow learning process in the middle layers
V-Net	3D CNN	Segmentation	Suitable for volumetric data	High computational costs
RNN, LSTM, GRU	RNN	Sequential tasks, Segmentation	Models the time dependence	Gradient vanishing problem
R-CNN	CNN + RPN	Segmentation, Object Detection	Enhance the precision of the region proposal and object distinction	Can be overloaded for simple tasks, Inability to be implemented in real time
DeepLab	CNN + CRF	Segmentation, Contours, etc	Improvement of borders thanks to CRFs	High complexity, can be slow
Transformer-based Models	Transformers	Sequential tasks, Computer Vision Tasks	Excellent ability to capture long-distance dependencies	High data and computing power requirements

6 Network Training Methods

In this section, we will discuss some of the most notable training techniques employed in the training of MIS networks.

1. **Deeply Supervised:** Direct supervision is given to the hidden layers in deep supervision, and this is also propagated to the lower layers. The idea was implemented in [73] and operationalized using the companion objective function, which GoogLeNet and other well-known networks used in the hidden layers [13].
2. **Weakly Supervised:** In supervised approaches for MIS, the requirement for pixel-level annotation is often unmet, and even when it is available, it is often a tedious and expensive process. One potential solution to this problem is to utilize image-labeled data, for instance, a binary label indicating the presence or absence of a pattern. This approach is exemplified in [74], where "point labels" are employed to indicate the presence of a nodule, thereby reducing the system's reliance on fully annotated images. This approach can be considered a weakly supervised learning technique [13].
3. **Transfer Learning:** Transfer learning is the capacity of a system to recognize and utilize previously acquired knowledge. This can be achieved in two distinct ways: firstly, by fine-tuning a pre-trained network on general images, and secondly, by fine-tuning a network that has been pre-trained on medical images of a different target organ. This is the most widely used method, and it has been demonstrated to yield superior performance when the source and target network tasks are analogous. Even transferring weights is more effective than random initialization [13].

7 Challenges and Future Prospects

One of the biggest challenges in MIS is data, as DL models rely heavily on large annotated datasets. However, medical image annotation has always been a difficult task, as it requires expertise in the medical field, making it difficult to find qualified annotators, and privacy issues make it difficult to obtain large amounts of data [75].

For a medical image dataset, the image modality affects how it is processed and changes the used model. Since we may find multi-modality data, it is difficult and challenging to adapt the model to it [75].

Medical image data often contains classes that appear more than others, called the class imbalance problem, which leads to great difficulty in training the DL model and makes the model's accuracy misleading [13].

Human organs vary in size, shape, and location, and the heterogeneous appearance of the organ to be segmented is one of the major challenges in segmenting medical images [13].

The problems presented previously can be partially solved with the data augmentation technique, as it allows us to create new data, as well as control the

distribution of classes in it, as well as using the weakly supervised learning technique to train models, which reduces the need to annotate the data.

Regarding the future research direction, we suggest focusing on the transformer model for domain adaptation, as well as attention techniques, to solve the multi-modality model adaptation problem. We also suggest delving into data augmentation techniques, especially the Generative Adversarial Networks (GANs), and trying to integrate attention into them to improve the quality and accuracy of the generated images.

8 Conclusion

The field of disease diagnosis using DL techniques is currently experiencing a period of rapid progress, driven by the current prevalence of diseases and the importance of reducing the burden on doctors. This has prompted us to provide an overview of some of the most widely used models, accompanied by an analysis of the challenges that have led to the emergence of technologies that are more advanced than their predecessors. In addition, we have included a discussion of the evaluation methods used in these models, as well as an overview of the types of medical data and their sources, the main challenges facing the field, and a suggestion for future research directions. We focused on the most important technology, namely transformers and attention techniques which can still be developed.

References

1. C. Peng, X. Zhang, G. Yu, G. Luo, J. Sun, Large kernel matters - improve semantic segmentation by global convolutional network, CoRR abs/1703.02719 (2017). arXiv:1703.02719. URL <http://arxiv.org/abs/1703.02719>
2. P. Malhotra, S. Gupta, D. Koundal, A. Zaguia, W. Enbeyle, Deep neural networks for medical image segmentation, *Journal of Healthcare Engineering* 2022 (2022).
3. Lawrence G Roberts. Machine perception of three-dimensional solids. PhD thesis, Massachusetts Institute of Technology, 1963.
4. K.S. Fu and J.K. Mui. A survey on image segmentation. *Pattern Recognition*, 13(1):3–16, 1981. ISSN 0031-3203. doi: [https://doi.org/10.1016/0031-3203\(81\)90028-5](https://doi.org/10.1016/0031-3203(81)90028-5). URL <https://www.sciencedirect.com/science/article/pii/0031320381900285>.
5. Yu-Jin Zhang. An overview of image and video segmentation in the last 40 years. *Advances in Image and Video Segmentation*, pages 1–16, 2006.
6. X. Liu, L. Song, S. Liu, Y. Zhang, A review of deep-learning-based medical image segmentation methods, *Sustainability* 13 (3) (2021). <https://doi.org/10.3390/su13031224>. URL <https://www.mdpi.com/2071-1050/13/3/1224>
7. Thoma, Martin. "A survey of semantic segmentation." arXiv preprint arXiv:1602.06541 (2016).
8. R. C. Gonzalez, R. E. Woods, *Digital Image Processing (3rd Edition)*, Prentice-Hall, Inc., USA, 2006.

9. H. G. Kaganami, Z. Beiji, Region-based segmentation versus edge detection, in: 2009 Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, 2009, pp. 1217–1221. <https://doi.org/10.1109/IIH-MSP.2009.13>.
10. W.-X. Kang, Q.-Q. Yang, R.-P. Liang, The comparative research on image segmentation algorithms, in: 2009 First International Workshop on Education Technology and Computer Science, Vol. 2, 2009, pp. 703–707. <https://doi.org/10.1109/ETCS.2009.417>.
11. W. Qiaoping, One image segmentation technique based on wavelet analysis in the context of texture, *Data Collection and Processing* 13 (1) (1998) 12–16.
12. M. J. Islam, S. Basalamah, M. Ahmadi, M. A. Sid-Ahmed, Capsule image segmentation in pharmaceutical applications using edge-based techniques, in: 2011 IEEE INTERNATIONAL CONFERENCE ON ELECTRO/INFORMATION TECHNOLOGY, 2011, pp. 1–5. <https://doi.org/10.1109/EIT.2011.5978613>.
13. M. H. Hesamian, W. Jia, X. He, P. Kennedy, Deep learning techniques for medical image segmentation: achievements and challenges, *Journal of digital imaging* 32 (2019) 582–596.
14. F. Lateef, Y. Ruichek, Survey on semantic segmentation using deep learning techniques, *Neurocomputing* 338 (2019) 321–348. <https://doi.org/https://doi.org/10.1016/j.neucom.2019.02.003>. URL <https://www.sciencedirect.com/science/article/pii/S092523121930181X>.
15. J. Moorthy, U. D. Gandhi, A survey on medical image segmentation based on deep learning techniques, *Big Data and Cognitive Computing* 6 (4) (2022). <https://doi.org/10.3390/bdcc6040117>. URL <https://www.mdpi.com/2504-2289/6/4/117>.
16. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Vol. 30, Curran Associates, Inc., 2017. URL <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>.
17. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, OpenReview.net, 2021. URL <https://openreview.net/forum?id=YicbFdNTTy>.
18. E. U. Henry, O. Emebob, C. A. Omonhinmin, Vision transformers in medical imaging: A review (2022). <https://doi.org/10.48550/ARXIV.2211.10043>. URL <https://arxiv.org/abs/2211.10043>.
19. J. Li, J. Chen, Y. Tang, C. Wang, B. A. Landman, S. K. Zhou, Transforming medical imaging with transformers? a comparative review of key properties, current progresses, and future perspectives, *Medical Image Analysis* 85 (2023) 102762. <https://doi.org/https://doi.org/10.1016/j.media.2023.102762>. URL <https://www.sciencedirect.com/science/article/pii/S1361841523000233>.
20. F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, H. Fu, Transformers in medical imaging: A survey (2022). <https://doi.org/10.48550/ARXIV.2201.09873>. URL <https://arxiv.org/abs/2201.09873>.
21. J. Beutel, H. L. Kundel, Y. Kim, R. L. Van Metter, S. C. Horii, *Handbook of medical imaging: display and PACS*, Vol. 3, Spie Press, 2000.

22. A. Parvaiz, M. A. Khalid, R. Zafar, H. Ameer, M. Ali, M. M. Fraz, Vision transformers in medical computer vision – a contemplative retrospection (2022). <https://doi.org/10.48550/ARXIV.2203.15269>. URL <https://arxiv.org/abs/2203.15269>.
23. K. He, C. Gan, Z. Li, I. Rekik, Z. Yin, W. Ji, Y. Gao, Q. Wang, J. Zhang, D. Shen, Transformers in medical image analysis, *Intelligent Medicine* (2022). doi:<https://doi.org/10.1016/j.imed.2022.07.002>
24. S.-Y. Huang, W.-L. Hsu, R.-J. Hsu, D.-W. Liu, Fully convolutional network for the semantic segmentation of medical images: A survey, *Diagnostics* 12 (11) (2022). <https://doi.org/10.3390/diagnostics12112765>. URL <https://www.mdpi.com/2075-4418/12/11/2765>.
25. Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, *Proceedings of the IEEE* 86 (11) (1998) 2278–2324. <https://doi.org/10.1109/5.726791>.
26. A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: F. Pereira, C. Burges, L. Bottou, K. Weinberger (Eds.), *Advances in Neural Information Processing Systems*, Vol. 25, Curran Associates, Inc., 2012. URL <https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf>.
27. M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), *Computer Vision – ECCV 2014*, Springer International Publishing, Cham, 2014, pp. 818–833.
28. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA, 2016, pp. 770–778. <https://doi.org/10.1109/CVPR.2016.90>. URL <https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.90>.
29. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, *Computational and Biological Learning Society*, 2015, pp. 1–14.
30. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9. <https://doi.org/10.1109/CVPR.2015.7298594>.
31. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: 2016 IEEE Conference on 22 Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2818–2826. <https://doi.org/10.1109/CVPR.2016.308>.
32. C. Szegedy, S. Ioffe, V. Vanhoucke, A. A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI’17*, AAAI Press, 2017, p. 4278–4284.
33. F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA, 2017, pp. 1800–1807. <https://doi.org/10.1109/CVPR.2017.195>. URL <https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.195>.
34. F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, K. Keutzer, Squeezenet: Alexnet-level accuracy with 50x fewer parameters and 10.5mb model size (2016). <https://doi.org/10.48550/ARXIV.1602.07360>. URL <https://arxiv.org/abs/1602.07360>.

35. G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2261–2269. <https://doi.org/10.1109/CVPR.2017.243>.
36. W. Zhang, R. Li, H. Deng, L. Wang, W. Lin, S. Ji, D. Shen, Deep convolutional neural networks for multi-modality isointense infant brain image segmentation, *NeuroImage* 108 (2015) 214–224. <https://doi.org/https://doi.org/10.1016/j.neuroimage.2014.12.061>. URL <https://www.sciencedirect.com/science/article/pii/S1053811914010660>.
37. A. Prasoon, K. Petersen, C. Igel, F. Lauze, E. Dam, M. Nielsen, Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network, in: K. Mori, I. Sakuma, Y. Sato, C. Barillot, N. Navab (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 246–253.
38. P. Moeskops, J. M. Wolterink, B. H. M. van der Velden, K. G. A. Gilhuijs, T. Leiner, M. A. Viergever, I. Išgum, Deep learning for multi-task medical image segmentation in multiple modalities, in: S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, W. Wells (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, Springer International Publishing, Cham, 2016, pp. 478–486.
39. J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3431–3440. <https://doi.org/10.1109/CVPR.2015.7298965>.
40. D. Nie, L. Wang, Y. Gao, D. Shen, Fully convolutional networks for multi-modality isointense infant brain image segmentation, in: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), 2016, pp. 1342–1345. <https://doi.org/10.1109/ISBI.2016.7493515>.
41. H. R. Roth, H. Oda, Y. Hayashi, M. Oda, N. Shimizu, M. Fujiwara, K. Misawa, K. Mori, Hierarchical 3d fully convolutional networks for multi-organ segmentation (2017). <https://doi.org/10.48550/ARXIV.1704.06382>. URL <https://arxiv.org/abs/1704.06382>.
42. Q. Wang, Q. Liu, G. Luo, Z. Liu, J. Huang, Y. Zhou, Y. Zhou, W. Xu, J.-Z. Cheng, Automated segmentation and diagnosis of pneumothorax on chest x-rays with fully convolutional multi-scale scse-densenet: a retrospective study, *BMC Medical Informatics and Decision Making* 20 (14) (2020) 1–12.
43. O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: N. Navab, J. Hornegger, W. M. Wells, A. F. Frangi (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Springer International Publishing, Cham, 2015, pp. 234–241.
44. R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, A. K. Nandi, Medical image segmentation using deep learning: A survey, *IET Image Processing* 16 (5) (2022) 1243–1267. <https://doi.org/10.1049/ipr2.12419>. URL <https://doi.org/10.1049/2Fipr2.12419>.
45. Y. Gordienko, P. Gang, J. Hui, W. Zeng, Y. Kochura, O. Alienin, O. Rokovy, S. Stirenko, Deep learning with lung segmentation and bone shadow exclusion techniques for chest x-ray analysis of lung cancer, in: Z. Hu, S. Petoukhov, I. Dychka, M. He (Eds.), *Advances in Computer Science for Engineering and Education*, Springer International Publishing, Cham, 2019, pp. 638–647.
46. Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, O. Ronneberger, 3d u-net: Learning dense volumetric segmentation from sparse annotation, in: S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, W. Wells (Eds.), *Medical Image Computing and*

- Computer-Assisted Intervention – MICCAI 2016, Springer International Publishing, Cham, 2016, pp. 424–432.
47. H. Cui, X. Liu, N. Huang, Pulmonary vessel segmentation based on orthogonal fused u-net++ of chest ct images, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI, Springer-Verlag, Berlin, Heidelberg, 2019, p. 293–300. https://doi.org/10.1007/978-3-030-32226-7_33. URL https://doi.org/10.1007/978-3-030-32226-7_33.
 48. Q. Jin, Z. Meng, C. Sun, H. Cui, R. Su, Ra-unet: A hybrid deep attention-aware network to extract liver and tumor in ct scans, *Frontiers in Bioengineering and Biotechnology* 8 (2020). <https://doi.org/10.3389/fbioe.2020.605132>. URL <https://www.frontiersin.org/articles/10.3389/fbioe.2020.605132>.
 49. C. Guo, M. Szemenyei, Y. Pei, Y. Yi, W. Zhou, Sd-unet: A structured dropout u-net for retinal vessel segmentation, in: 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE), 2019, pp. 439–444. <https://doi.org/10.1109/BIBE.2019.00085>.
 50. F. Milletari, N. Navab, S.-A. Ahmadi, V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: 2016 Fourth International Conference on 3D Vision (3DV), 2016, pp. 565–571. <https://doi.org/10.1109/3DV.2016.79>.
 51. E. Gibson, F. Giganti, Y. Hu, E. Bonmati, S. Bandula, K. Gurusamy, B. Davidson, S. P. Pereira, M. J. Clarkson, D. C. Barratt, Automatic multi-organ segmentation on abdominal ct with dense v-networks, *IEEE Transactions on Medical Imaging* 37 (8) (2018) 1822–1834. <https://doi.org/10.1109/TMI.2018.2806309>.
 52. M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, V. K. Asari, Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation (2018). <https://doi.org/10.48550/ARXIV.1802.06955>. URL <https://arxiv.org/abs/1802.06955>.
 53. Y. Gao, J. M. Phillips, Y. Zheng, R. Min, P. T. Fletcher, G. Gerig, Fully convolutional structured lstm networks for joint 4d medical image segmentation, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), 2018, pp. 1104–1108. <https://doi.org/10.1109/ISBI.2018.8363764>.
 54. J. Cai, L. Lu, Y. Xie, F. Xing, L. Yang, Improving deep pancreas segmentation in ct and mri images via recurrent neural contextual learning and direct loss function (2017). <https://doi.org/10.48550/ARXIV.1707.04912>. URL <https://arxiv.org/abs/1707.04912>.
 55. R. P. K. Poudel, P. Lamata, G. Montana, Recurrent fully convolutional neural networks for multi-slice mri cardiac segmentation, in: M. A. Zuluaga, K. Bhatia, B. Kainz, M. H. Moghari, D. F. Pace (Eds.), *Reconstruction, Segmentation, and Analysis of Medical Images*, Springer International Publishing, Cham, 2017, pp. 83–94.
 56. J. Koutnik, K. Greff, F. Gomez, J. Schmidhuber, A clockwork rnn, in: E. P. Xing, T. Jebara (Eds.), *Proceedings of the 31st International Conference on Machine Learning*, Vol. 32 of *Proceedings of Machine Learning Research*, PMLR, Beijing, China, 2014, pp. 1863–1871. URL <https://proceedings.mlr.press/v32/koutnik14.html>.
 57. R. Girshick, Fast r-cnn, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1440–1448. <https://doi.org/10.1109/ICCV.2015.169>.
 58. S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (6) (2017) 1137–1149. <https://doi.org/10.1109/TPAMI.2016.2577031>.

59. K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988. <https://doi.org/10.1109/ICCV.2017.322>.
60. L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, Semantic image segmentation with deep convolutional nets and fully connected crfs (2014). <https://doi.org/10.48550/ARXIV.1412.7062>. URL <https://arxiv.org/abs/1412.7062>.
61. L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40 (4) (2018) 834–848. <https://doi.org/10.1109/TPAMI.2017.2699184>.
62. L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation (2017). <https://doi.org/10.48550/ARXIV.1706.05587>. URL <https://arxiv.org/abs/1706.05587>.
63. L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.), *Computer Vision – ECCV 2018*, Springer International Publishing, Cham, 2018, pp. 833–851.
64. W. Tang, D. Zou, S. Yang, J. Shi, Dsl: Automatic liver segmentation with faster r-cnn and deeplab, in: V. Kůrková, Y. Manolopoulos, B. Hammer, L. Iliadis, I. Maglogiannis (Eds.), *Artificial Neural Networks and Machine Learning – ICANN 2018*, Springer International Publishing, Cham, 2018, pp. 137–147.
65. J. Wang, X. Liu, Medical image recognition and segmentation of pathological slices of gastric cancer based on deeplab v3+ neural network, *Computer Methods and Programs in Biomedicine* 207 (2021) 106210. <https://doi.org/https://doi.org/10.1016/j.cmpb.2021.106210>. URL <https://www.sciencedirect.com/science/article/pii/S0169260721002844>.
66. L. Ahmed, M. M. Iqbal, H. Aldabbas, S. Khalid, Y. Saleem, S. Saeed, Images data practices for semantic segmentation of breast cancer using deep neural network, *Journal of Ambient Intelligence and Humanized Computing* (2020) 1–17.
67. J. Wang, L. Wei, L. Wang, Q. Zhou, L. Zhu, J. Qin, Boundary-aware transformers for skin lesion segmentation, in: M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, C. Essert (Eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, Springer International Publishing, Cham, 2021, pp. 206–216.
68. K. Deng, Y. Meng, D. Gao, J. Bridge, Y. Shen, G. Lip, Y. Zhao, Y. Zheng, Transbridge: A lightweight transformer for left ventricle segmentation in echocardiography, in: *Simplifying Medical Ultrasound: Second International Workshop, ASMUS 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings*, Springer-Verlag, Berlin, Heidelberg, 2021, p. 63–72. https://doi.org/10.1007/978-3-030-87583-1_7. URL https://doi.org/10.1007/978-3-030-87583-1_7.
69. Z. Shen, H. Yang, Z. Zhang, S. Zheng, Automated kidney tumor segmentation with convolution and transformer network, in: N. Heller, F. Isensee, D. Trofimova, R. Tejpaul, N. Papanikolopoulos, C. Weight (Eds.), *Kidney and Kidney Tumor Segmentation*, Springer International Publishing, Cham, 2022, pp. 1–12.
70. W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, J. Li, Transbts: Multimodal brain tumor segmentation using transformer, in: M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, C. Essert (Eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, Springer International Publishing, Cham, 2021, pp. 109–119.

71. Q. Jia, H. Shu, Bitr-unet: A cnn-transformer combined network for mri brain tumor segmentation, in: A. Crimi, S. Bakas (Eds.), *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Springer International Publishing, Cham, 2022, pp. 3–14.
72. D. Karimi, S. D. Vasylechko, A. Gholipour, Convolution-free medical image segmentation using transformers, in: M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, C. Essert (Eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, Springer International Publishing, Cham, 2021, pp. 78–88.
73. C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, Z. Tu, Deeply-Supervised Nets, in: G. Lebanon, S. V. N. Vishwanathan (Eds.), *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, Vol. 38 of *Proceedings of Machine Learning Research*, PMLR, San Diego, California, USA, 2015, pp. 562–570. URL <https://proceedings.mlr.press/v38/lee15a.html>.
74. R. Anirudh, J. J. Thiagarajan, T. Bremer, H. Kim, Lung nodule detection using 3D convolutional neural networks trained on weakly labeled data, in: G. D. Tourassi, S. G. Armato (Eds.), *Medical Imaging 2016: Computer- Aided Diagnosis*, Vol. 9785 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, 2016, p. 978532. <https://doi.org/10.1117/12.2214876>.
75. Li, Yuemeng, and Yong Fan. "Medical Image Segmentation with Domain Adaptation: A Survey." arXiv preprint arXiv:2311.01702 (2023).

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

