**Università degli Studi di Modena e Reggio Emilia**

DIPARTIMENTO DI INGEGNERIA "ENZO FERRARI"

CORSO DI LAUREA MAGISTRALE IN INGEGNERIA INFORMATICA

# Inferior Alveolar Canal Segmentation using Deep Neural Networks

*Relatore:*
Prof. Costantino Grana

*Correlatore:*
Prof. Federico Bolelli

*Candidato:*
Luca Lumetti

ANNO ACCADEMICO 2021-2022

# *Abstract*

**Inferior Alveolar Canal Segmentation using Deep Neural Networks**

**Keywords:** Medical Imaging, Cone Bean Computed Tomography, Inferior Alveolar Canal, Image Segmentation, Deep Learning.

# *Abstract in Italian*

**Utilizzo di Reti Neurali per la Segmentazione del Canale Alveolare Inferiore**

**Parole Chiave:** Medical Imaging, Cone Bean Computed Tomography, Inferior Alveolar Canal, Image Segmentation, Deep Learning.

# *Summary in Italian*

Il regolamento della Facoltà di Ingegneria di Modena prevede che le tesi scritte in lingua inglese debbano contenere un *abstract* ed un'ampia sintesi dei contenuti in lingua italiana: in accordo a questa regola, proponiamo un sunto degli argomenti, delle tecniche e dei risultati che verranno delineati nell'elaborato. Si noti che non si tratta di un sunto esaustivo, e che non è possibile valutare la tesi dalla semplice lettura di queste righe. Per una descrizione più dettagliata e più rigorosa, e per i risultati sperimentali ottenuti, si rimanda al testo in inglese.

Questa tesi tratta del... Durante il lavoro svolto è inoltre stato pubblicato un *paper* attualmente in fase di revisione (si veda l'Appendice per ulteriori dettagli).

*To ...*

# *Acknowledgements*

Foremost, I would like to express my sincere gratitude to ...

# Contents

# List of Figures

# List of Tables

# List of Listings

# Chapter 1

# Segmentation of the Inferior Alveolar Canal: an overview

## 1.1 Introduction

Dental implant placement within the jawbone is a routine surgical procedure that can become complicated due to the presence of the Inferior Alveolar Nerve (IAN) nearby. The nerve, in particular, is frequently in close proximity to the roots of molars, and its position must thus be meticulously detailed prior to surgical removal. Avoiding contact with the IAN is a primary concern during these operations, thus its segmentation is crucial in surgical planning. Today the standard de-facto is to take a CBCT scan of the jawbone and a 2D panomaric view is extracted. This view allow medical experts to depict the IANs position with line. We refer to this type of annotation as *sparse* or *2D* annotation. A 3D annotation of the IAC is often avoided as it would require a huge amount of time, but this type of segmentation would offer a much precise knowledge of the position of the IAN and IAC and could allow dentists to plan a more detailed surgical approach. For this reason a lot of research about automatic segmentation of the IAC has been carried out and is still active today.

In this chapter we first describe in details the role of the IAN and the IAC, what a CBCT is and the definitions and carachteristics of different types of segmentations, to give a brief introduction on which are the main components of the final work. In the following chapter we will look at the main components used to perform image segmentation in the medical field, until to today state of the art. Next in chapter 3 we will detail the dataset that has been used, by describing how it has been created, preprocessed and some other thoughts. In chapter 4, the network I've used as starting point and the benefits obtained by refactoring the code. Finally in chapter 5 we will discuss the different approach taken to try to improve the current state of the art, the results obtained and some of the possible future works.

## 1.2 Inferior Alveolar Canal

The Inferior Alveolar Canal (IAC) is a small passageway shaped as a tube that runs through the lower jawbone. It houses the Inferior Alveolar Nerve (IAN), which is responsible for transmitting sensory information from the teeth, gums and lips to the brain. It also provides motor innervation to the muscles of mastication (i.e. the muscles responsible for chewing). Dentists need to be able to accurately locate the IAC before performing certain surgical operations, such as tooth extractions or placement of dental implants. This is because the IAC is located very close to the roots of the teeth, and damage to the IAC during surgery can result in permanent nerve damage which would cause numbness, tingling, and pain in the affected area. In severe cases, it can also lead to muscle weakness and paralysis.

## 1.3 Cone Beam Computed Tomography

Cone beam computed tomography (CBCT) is a medical imaging technique consisting of X-ray computed tomography where rays are divergent, forming a cone. This type of computed tomography is well suited for imaging the craniofacial area as it provides clear images of highly contrasted structures, very helpfull to evaluate bones. It has become common in dentistry such as oral surgery, endodontics and orthodontics. The main reasons and advantages of CBCT with respect to other CTs are:

1. **X-ray beam limitation:** reducing the size of the irradiated area by collimating the primary x-ray beam to the are of interest minimizes the radiation dose. Most CBCT units can be adjusted to scan small regions for specific diagnostic task. They are also able to scan the whole craniofacial structure if needed.

2. **Image accuracy:** We created a novel, large, and publicly available maxillo-facial CBCT (Cone Beam Computed Tomography) dataset, with 2D and 3D manual annotations, provided by expert clinicians. All CBCT units provide voxel resolutions that are isotropic (i.e. equals in all the 3 dimensions) while in conventional CT, voxel are anisotripic (i.e. rectangular cubes).

3. **Rapid scan time:** CBCT acquires all the basis images in a single rotation, thus scan time goes from 10s to 70s. Although faster scanning time usually means fewer basis images from which to reconstruct the volumetric dataset, motion artifacts due to subject movement are reduced.

These advantages come with some drawbacks: Hounsfield units (HU) is the metric used to determine the radiodensity of tissue analized. In the Hounsfield scale, numbers go from values of $-1000$ for air to values of $1600$ for dense bones. In CBCT scans, the radiodensity is inaccurate because different areas in the scan appear with different greyscale values depending on their relative positions in the organ being scanned, despite possessing identical densities, because the image value of a voxel of an organ depends on the position in the image volume. HU measured from the same anatomical area with both CBCT and medical-grade CT scanners are not identical and are thus unreliable for determination of site-specific, radiographically-identified bone density for purposes such as the placement of dental implants, as there is ”no good data to relate the CBCT HU values to bone quality” [1].
The images resulting from a CBCT scans are usually exported as DICOM (Digital Imaging and Communications in Medicine) which is the standard used worldwide to store, exchange, and transmit medical images.

## 1.4 DICOM file format

TODO?

## 1.5 Image Segmentation

Image segmentation is a well known topic in computer and image processing with a wide range of application, such as medical imaging, robotics, video surveillance, etc. It involves partitioning images into one or more objects and can also includes classify these objects.

Many traditional algorithms have been developed in the literature but, in the most recent years, they have all been dominated by deep neural networks. Since the 2015 a huge amount of different types of networks that aim to perform image segmentation has been proposed for each of the field where it's needed. Before presenting how nowday segmentation is performed, we must state which are the different type of segmentation that have been classified:

- **Semantic segmentation:** Semantic Segmentation perform a pixel-by-pixel classification with a predefined set of objects categories for all the pixels of the images. In pratice, given a RGB image (`height` × `width` × `3`) we output a segmentation map of size (`height` × `width` × `classes`) where each value correspond to which class the same pixel in the original images belongs.
- **Instance segmentation:** One possible issue of semantic segmentation is that it doesn't allow to distinguish two or more object of the same class when they overlap in the image. Instance segmentation overcome this problem by outputting a different number of channels based on the number of instances present in the image.
- **Panoptic segmentation:** The latter type of segmentation is called Panoptic segmentation and is the result of the previously presented method joined together. The difference with the instance segmentation is that in this case instances are not allowed to overlap then for to a single pixel, a single instance must be assigned.

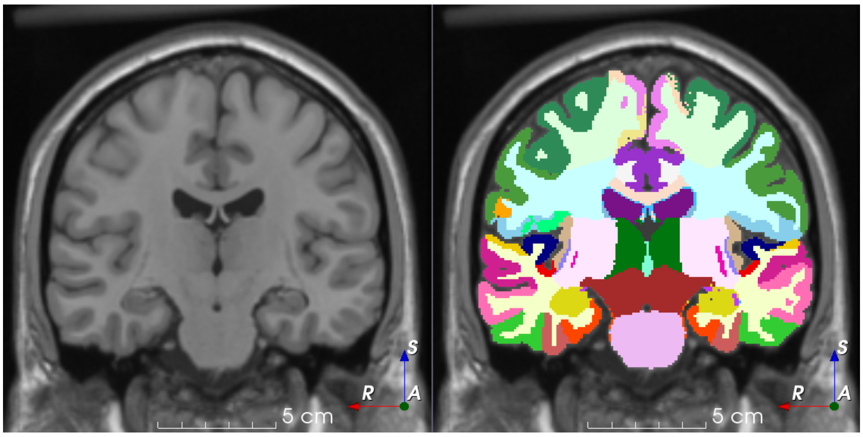

FIGURE 1.1: Example of a multiclass semantic segmentation, different colors represent different classes

# Chapter 2

# Segmentation Neural Networks

## 2.1 Supervised Learning

In the medical image segmentation tasks, supervised learning is the most popular method, where the latests improvements mainly includes network backbones, network blocks and the design of novel loss functions.

### 2.1.1 Backbone networks

Encoder-decoder structure is the most popular end-to-end architectures, such fully convolutional networks (FNC) like U-Net, Deeplab, and SegNet. The encoder part aim to extract high level features from the input image and project them into a latent space. The decoder part instead restore the extracted features from the latent space to the original space size.

#### 2.1.1.1 U-Net and V-Net

In 2015, Ronneberger et al. proposed U-Net which has been widely used for medical image segmentation and many variants have been proposed also in the recent years. The stucture of U-Net is symmetrical and the main scope is to fuse low-level features, which medical images are usually noisy but show blurred boundaries, with high level features via skip connections.

U-Net was designed to deal with 2D images but when dealing with medical images, we usually have to handle 3D images. A straighforward solution was to extract 2D slices from the original image, fed them to the network and stack the outputs to obtain the final 3D output. The main drawback of this approach is that we lost the spatial information among the slices as they are threated as independent images. Motivated by this idea, Çiçek et al. proposed a solution to this problem by using 3D convolutions inside U-Net. This network, named 3D U-Net, includes only three down-sampling steps because of the high computational cost of 3D Convolutions, leading to a less effective extraction of of deep-layer image features. Milletari et al. proposed a similar architecture, named V-Net, which emply more skip connections than U-Net to design a deeper network. Some Recurrent Neural Network mixed with U-Net has been proposed to model the time dependence of image sequences.

#### 2.1.1.2 Recurrent Neural Networks

Gao et al. joined LSTM and CNN to model the temporal relationships between different brain MRI slices to improve segmentation accuracy. Clearly, RNN can capture local and global spatial features of images by considering the context information relationship. However, in medical image segmentation, the capture of complete and valid temporal information requires good medical image quality (e.g. smaller slice thickness and pixel spacing). Therefore, the design of RNN is uncommon for improving the performance of medical image segmentation.

#### 2.1.1.3 Pseudo-3D Networks

As stated before, most of the medical image are 3D images, but using 3D convolution requires a lot of computational resources. Therefore some pseudo-3D segmentation methods have been proposed. For example, Vu et al. applied the overlay of adjacent slices as input to the centeral slice prediction and then fed the obtained 2D feature map into a standar 2D network.

#### 2.1.1.4 Generative Adversarial Networks

Another type of networks that have been exploited in medical image segmentation are GANs, mostly for data augmentation by generating new samples. Pollastri et al. remodelled two different weel known GANs, Deep Convolutional GAN and a Laplacian GAN, to generate both skin lesion images and their segmentation masks, making the augmentation process extremely straighforward. In addition, the incorporation of the prior knowledge about organ shape and position may be crucial for improving medical image segmentation effect, where images are corrupted and thus contain artefacts due to limitations of imaging techniques. However, there are few works about how to incorporate prior knowledge into CNN models. As one of the earliest studies in this field, Oktay et al. proposed a novel and general method to combine a priori knowledge of shape and label structure into the anatomically constrained neural networks (ACNN) for medical image analysis tasks. In this way, the neural network training process can be constrained and guided to make more anatomical and meaningful predictions, especially in cases where input image data is not sufficiently informative or consistent enough (e.g., missing object boundaries).

After proposing U-Net in [7], the encoder-decoder structure became the most popular structure in medical image segmentation. The design of the network backbone focuses on more efficient feature extraction in the encoder and feature recovery and fusion in the decoder to improve segmentation accuracy.

### 2.1.2 Network Function Block

#### 2.1.2.1 Dense connection

Dense connection is the most popular network block in medical image segmentation, used to contruct a kin of special convolution neural networks. The input of each layer comes from the output of all previous layers in the process of forward transmission. Inspired by this design, Guan et al. proposed an improved U-Net by replacing each sub-block of U-Net with a form of dense connection. Although the dense connection is helpful for obtaining riher image fetures, it often reduces the robustness of feature representation to a certian extent and increase the number of parameters.

#### 2.1.2.2 Inception block

For CNNs, deep networks often given better performances than shallow ones, buyt they encountr some new problems such as vanishing gradient, high memory usage, and slow convergence. The inception structure used in GoogleNet overcome this problems, and for this reason it has been also used over medical images. Gu et al. proposed CE-Net by introducing the inception structure and atrous convolution to each parallel structure to extrtact features on a wide reception field. Such complex structure however lead to a difficult model modification.

#### 2.1.2.3 Depth separability

To reduce the computational cost of 3D convolutions and their memory usage, Howard et al. proposed MobileNet to decompose vanilla convolutions into depthwise separable convolution and pointwise convolution.

#### 2.1.2.4 Attention

Attention block can selectively change input or asssigns different weights to input variables according to different importance.
*Spatial Attention* block aims to calculate the feature importance of each pixel in space-domain and extract the key information of an image. Oktay et al. proposed attention U-Net, where attention blocks were used to change the output of the encoder before fusing features from the encoder and the corresponding decoder. The attention block outputs a gating signal to control feature importance of pixels at different spatial positions.
Another type of attention block is the *Channel attention*, which utilizes learned global information to emphasize selectively useful features and suppress useless features. Hu et al. proposed SE-Net that introduced the channel attention to the field of image analysis and won the ImageNet Challenge in 2017.
Spatial and channel attention mechanisms are the two most popular strategies for improving feature representation. However, spatial attention ignores the difference of different channel information and treats each channel equally. On the contrary, the channel attention pools global information directly while ignoring local information in each channel, which is a relatively rough operation. Therefore, combining advantages of two attention mechanisms, researchers have designed many models based on a *mixed domain attention block*.

### 2.1.3 Loss functions

In addition to improved segmentation speed and accuracy by designing network backbone and the function block, designing new loss functions also resulted in improvements in subsequent inference-time segmentation accuracy. Therefore, a great deal of work has been reported about the design of suitable loss functions for medical image segmentation tasks.

#### 2.1.3.1 Cross Entropy

The cross entropy loss has been the most popular loss function. It compares pixel-wisely the predicted category vector with the real segmentation result vector and is defined as:

$$L_{CE} = -\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij} \log \hat{y}_{ij}$$

where $y_{ij}$ is the real segmentation result vector, $\hat{y}_{ij}$ is the predicted segmentation result vector, $n$ is the number of pixels in the image, and $C$ is the number of categories. The cross entropy loss is easy to implement and has been widely used in medical image segmentation tasks. However, it is not suitable for segmentation tasks with imbalanced data, because it does not consider the class imbalance problem. Therefore, a *weighted cross entropy loss* and *balanced cross entropy* have been proposed to solve this problem, where a $\beta$ hyperparameters is added to the loss function to adjust the weight of each class.

#### 2.1.3.2 Dice Loss

The Dice coefficient is a popular metric for the evaluation of medical image segmentation tasks. This metric is the measure of overlap between a segmentation result and its corresponding ground thruth:

$$DSC = \frac{2 \times |A \cap B|}{|A| + |B|}$$

where $A$ and $B$ are the segmentation result and the ground truth, respectively. The *Dice Loss* is then formulted as follow:

$$L_{DSC} = 1 - \frac{2 \times y \times \hat{y} + 1}{y + \hat{y} + 1}$$

Here 1 is added to the denominator to avoid the division by zero it the edge case when both $y$ and $\hat{y}$ are zero. The Dice loss is a good choice even for uneven samples, however it easly influences the back propagation and leads to a tranining difficulty.

#### 2.1.3.3 Generalized Dice Loss

Although the Dice Loss can solve the problem of class imbalance to a certain extent, it does not work for serius class imbalance. To solve this problem, researchers have proposed a *Generalized Dice Loss* that can be used for both binary and multi-class segmentation tasks. The generalized Dice loss is defined as:

$$L_{GDSC} = 1 - \frac{1}{m}\frac{2\sum_{j=1}^{m}\omega_j\sum_{i=1}^{n} y_{ij}\hat{y}_{ij}}{\sum_{j=1}^{m}\omega_j\sum_{i=1} n(y_{ij} + \hat{y}_{ij})}$$

where the weight $\omega_j$ is used to adjust the weight of each class, and $\omega_j = 1/(\sum_{i=1}^{n} p_{ij})^2$.

#### 2.1.3.4 Boundary Loss

Another approach to solve the problem of class imbalance have been proposed by Kervadec et al. for the task of brain lesion segmentation. They proposed a Boundary Loss which aims to minimize the distance between segmented boundaries and labeled boundaries. Results show that the combination of the Dice loss and the boundary loss is superior to the single ones. The composite loss is defined as:

$$L = \alpha L_{DSC} + (1 - \alpha)L_B$$

where the Boundary Loss $L_B$ for a binary segmentation is defined as:

$$L_B = \sum \phi(y) \times \hat{y}$$

where $\phi(y)$ is the signed distance function applied to the real segmentation $y$.

For medical image segmentation, the improvement of loss mainly focuses on the problem of segmentation of small objects in a large background (the problem of class imbalance). Chen et al. proposed a new loss function by applying traditional active contour energy minimization to convolutional neural networks, Li et al. proposed a new regularization term to improve the crossentropy loss function, and Karimi et al. proposed a loss function based on

Hausdorff distance (HD). Besides, there are still a lot of works trying to deal with this problem by adding penalties to loss functions or changing the optimization strategy according to specific tasks.

In many medical image segmentation tasks, there are often only one or two targets in an image, and the pixel ratio of targets is sometimes small, which makes network training difficult. Therefore, to improve network training and segmentation accuracy, it is easier to focus on smaller targets by changing loss functions than to change the network structure. However, the design of loss functions is highly task-specific, so we need to analyze carefully task requirement, and then design reasonable and available loss functions.

## 2.2 Weakly Supervised and Unsupervised Segmentation

Although convolutional neural networks show goods performances for medical image segmentation, results seriously depend on high-quality labels. In fact, it is rare to build many datasets with many high-quality labels, especially in the field of medical image analysis, since data acquisition and labeling often incur high costs. Therefore, a lot of studies on incomplete or imperfect datasets are reported. Unsupervised learning is a very important approach to improve the performance of medical image segmentation. In this section, we will introduce the weakly supervised learning, a method which make use of unsupervised learning for unlabelled data in combination with supervised learning with labelled data.

### 2.2.1 Data Augmentation

Performing data augmentation is mandatory in absence of a largely labeled datasets, and is still considered a good practice even when we have enough data. However, new data generated with this method produce images that are highly correlated with the original images.

#### 2.2.1.1 Traditional methods

With traditional methods we refer to all the computer vision techniques such as adding/removing noise, change brightness, saturation, contrast, colors, and change the image layout with rotations, distortion, scaling, etc. These technique are still very used today, usually combining them together with random parameters. Some of these algorithms may require a non negligible computational cost thus usually performed before the training procedure.

#### 2.2.1.2 Conditional Generative Adversarial Networks

As already described in Section 2.1.1.4, GAN and its variants have been widely for data augmentation. In particular, cGANs are often used in combination with standard GANs to generate labels relative to a given synthetic image.

### 2.2.2 Transfer Learning

Pre-trained model's parameters are often used to initialize a new model, transfer learning can achieve fast training for data with limited labels. The most popular approach is to use a model pretrained on ImageNet before performing the training on the medical data. Experiments demonstrated that this approach is useful as it improves the accuracy of segmentations. However the domain adaptation may be a problem when applying models trained over natural images to medical images analysis tasks. Moreover, these methods are hardly applicable to 3D medical image analysis because such pre-trained model rely on 2D datasets.
Hatamizadeh et al. recently proposed an unsupervised approach to pre-train a given model by relying only on unannotated medical images of the same domain of the main tasks. In pratice, the network is trained to perform a variety of tasks such as image reconstruction, classification of the rotation applied to the original image, etc. which can be performed in an unsupervised manner. Later, these tasks heads of the network are detached and the training on the main task is performed. Such pretraining aim to train the network to learn how to extract high level feature from a specific type of medical images, such as CT or MRI.
Also Cipriano et al. recently proposed a pre-training approach based on sparse labels. This type of labels are way easier to obtain but are not as accurate as the real dense labels. They fed these labels to the network together with the input image for a given number of steps, then used the learned parameters to train the network to produce the segmentation by relying only on the input, without the sparse labels.

## 2.3 Current direction of research

Until now we described the most popular network structures and loss functions for medical image segmentation tasks that were proposed and used up to date. Since the raise in popularity of vision transformers and graph neural networks, some novelty architectures are being proposed in medical imaging. Results obtained are still not as good as those obtained with traditional architectures, aside some really specific tasks or datasets, but they are still interesting and promising.

### 2.3.1 Network Architecture Search

The design process of a network architecture is a very time consuming task, and it is often difficult to find the best architecture for a given task. Therefore, many researchers have proposed methods to automate the design of network architectures. Such methods, named NAS (Network Architecture Search), focus on the *search space*, *search strategy*, and *performance estimation*. The search space is a candidate collection of network structures to be searched. The search space is divided into a global search space that represents the search for the entire network structure, and a cell-based search space that searches only a few small structures that are assembled into a complete large network by the ways of stacking and stitching. The search strategy aims to find the optimal network structure as fast as possible in search spaces. Popular search strategies are often grouped into three categories, reinforcement-based learning, evolutionary algorithms, and gradients. Performance estimation strategy is the process of assessing how well the network structure performs on target datasets. For NAS techniques, researcher pay more attention to the improvement of search strategies since search space and performance estimation methods are usually rarely changed.

Isensee et al. argued that too much manual adjustment on network structure could lead to over-fitting for a given dataset, and therefore proposed a medical image segmentation framework no-newUNet (nnU-Net) that adapts itself to any new dataset. The nnUnet automatically adjusts all hyperparameters according to the properties of the given dataset without manual intervention. Therefore, the nnU-Net only relies on vanilla 2D UNet, 3D UNet, UNet cascade and a robust training scheme. It focuses on the stage of pre-processing (resampling and normalization), training (loss, optimizer settings, data augmentation), inference (patch-based strategies, test-time-augmentations integration, model integration, etc.), and post-processing (e.g., enhanced single pass domain). In practical applications, the improvements of network structure design usually depend on experiences without adequate interpretability theory support, Moreover, more complex network models indicate higher risk of over-fitting.

### 2.3.2 Graph Convolutional Neural Network

Graph Convolutional Neural Network (GCN) is a type of neural network that utilizes graph structure to process data. In practive the Euclidean space of the image can be converted into graphs that can be modeled using GCN.

Gao et al. designed a new graph pooling (gPool) and graph unpooling (gUnpool) operation based on GCN and proposed an encoder-decoder model namely graph U-Net. The graph U-Net achieves better performance than popular UNets by adding a small number of parameters. In contrast to traditional convolutional neural networks where deeper is better, the performance of the graph U-Net cannot be improved by increasing the depth of networks when the value of depth exceeds 4. However, the graph U-Net show stronger capability of feature encoding than popular U-Nets when the value of depth is smaller or equivalent to 4.

Yang et al. proposed the end-to-end conditional partial residual plot convolutional network CPR-GCN for automatic anatomical marking of coronary arteries. Authors showed that the GCN-based approach provided better performance and stronger robustness than traditional and recent depth learning based approaches. Results from these GCNs in medical image segmentations are promising, as the graph structure has high data representation efficiency and strong capability of feature encoding

### 2.3.3 Interpretable Shape Attentive Neural Network

Currently, many deep learning algorithms tend to make judgments by using ”memorized“ models that approximately fit to input data. As a result, these algorithms cannot be explained sufficiently and give convincing evidences for each specific prediction. Therefore, the study of the interpretability of deep neural networks is a hot topic at present. Sun et al. proposed the SAU-Net that focuses on the interpretability and the robustness of models. The proposed architecture attempts to address the problem of poor edge segmentation accuracy in medical images by using a secondary shape stream. Specially, the shape stream and the regular texture stream can capture rich shape-dependent information in parallel. Furthermore, both spatial and channel attention mechanism are used for the decoder to explain the learning capability of models at each resolution of U-Net. Finally, by extracting the learned shape and spatial attention maps, we can interpret the highly activated regions of each decoder block. The learned shape maps can be used to infer correct shapes of interesting categories learned by the model. The SAU-Net is able to learn robust shape features of objects via the gated shape stream, and is also more interpretable than previous works via built-in saliency maps using attention.

### 2.3.4 Vision Transformer

Recently, transformer-based architectures have become very popular that replaces the convolutional operator and use self-attention modules to compose entire encoderdecoder structures that can encode long-range dependencies. It has been a great success in the field of natural language processing. Dosovitskiy et al. proposed Vision Transformer (ViT) that is able to classify images directly using the Transformer. Recently, a large number of researches have applied the transformer to medical image segmentation. CNNs have a comparative advantage in extracting the underlying features. These low-level features form the key points, lines, and some basic image structures at the patch level. However, when we detect these basic visual elements, the higher-level visual semantic information is often more concerned with how these elements relate to each other to form an object, and how the spatial location of objects relates to each other to form the scene. At present, the transformer is more natural and effective in dealing with the relationships between these elements. However, if all the convolutional operators in CV tasks are replaced by Transformer, it may suffer from many problems, such as high computational cost and memory usage. From existing researches, the combination of Transformer and CNNs may lead to better results. Recently, Chen et al. proposed a U-Net shaped network, where the encoder was made of ViT only while the decoder was fully convolutional.

# Bibliography

[1] Dale Miles and Robert Danforth. A clinician's guide to understanding cone beam volumetric imaging (cbvi). *Acad Dent Ther Stomatol*, pages 1–13, 01 2007.