

# Medical Image Segmentation - How Deep is Your U-Net?

Giovani Tavares

giovanitavares20@gmail.com

Universidade of São Paulo — November 3, 2025

## 1 Motivation

I have been studying different neural network architectures in the Deep Learning course I am taking this semester as part of my Master's program in the University Of São Paulo. *U-Net* is one particular architecture that caught my attention because it is highly modular and flexible while still powerful and adaptable – a playground for neural network dissectors like me and maybe you. More importantly, *U-Nets* are widely used in medical imaging, helping to diagnose and treat many diseases, which make the contribution from their inventors even more commendable.

If you are interested in the interted in deep learning architectures used in medicine or is just curious to get a better grasp of *U-Nets* components, this post is for you.

In the following sections I will give a brief explanation of the *U-Net* architecture and then present an ablation study I did with it. If you are familiar with U-Nets you might already have some hint on what my ablation study was about and can jump right to section [?].

I used two popular public datasets for medical image segmentation, the International Skin Imaging Collaboration (ISIC) [4] , used for melanoma diagnosis; and DRIVE (Digital Retinal Images for Vessel Extraction) [1].

This post is divided in the following sections:

1. **U-Net Overview:** a brief explanation of the architecture. If you want to go deeper in the details I strongly recommend you to read the original paper: U-Net: Convolutional Networks for Biomedical Image Segmentation [3]
2. **U-Nets Ablation Study:** an explanation of the study I did to answer the question on what are the parts of U-Net that matter the most to the image segmentation task in the cited datasets
3. **Results:** the results of the U-Net dissect performed from a both task performance and technical overview.

So without further ado, let's go down to the experiment.

## 2 U-Net Overview

According to U-Net's introductory paper [3], this is an architecture that "consists of a contracting path to capture context and a symmetric expanding path that enables precise localization". Even though this is not a very detailed description of U-Net, it captures its essence, given that U-Nets are convolutional neural networks (CNNs) made up of two components: an encoder and a decoder that produce symmetric and opposite effects on their inputs. The **encoder** *contracts* its inputs by reducing its dimensions while the decoder expands its input's dimensions while

O objetivo do trabalho foi investigar os impactos do número de camadas de codificação e decodificação e das *skip connections* entre essas camadas na performance da rede no conjunto de dados **DRIVE: Digital Retinal Images for Vessel Extraction** a fim de responder à seguinte pergunta:

**Pergunta de pesquisa:** há algum *trade-off* entre as *skip connections* e o número de camadas de codificação/decodificação de **U-Nets**?

### 3 Metodologia

A fim de se responder à pergunta proposta, U-Nets de diferentes profundidades e com diferentes configurações de *skip connections* foram treinadas. Utilizou-se a biblioteca *PyTorch* para definir e treinar as redes.

O código foi base para definir a U-Net de segmentação foi o disponibilizado por Nicholas DiSalvo em [?]. Uma vez que [?] implementa a U-Net para construir um *Denoising Diffusion Probabilistic Model*, o código foi alterado pelo autor deste relatório para que a rede realizasse a tarefa de segmentação de imagens.

A U-Net implementada foi a mesma apresentada em [3], o artigo que introduziu essa arquitetura. Para isso, a imagem da arquitetura apresentada no artigo apoiou a definição das dimensões dos canais em cada camada da rede.

Já a função de perda utilizada durante o treinamento de todas as redes foi a *DiceBCELoss*, uma função que combina a *Dice* com a *Binary Cross-Entropy*. Ela foi implementada com base no artigo [2].

#### 3.1 Arquitetura U-Net

A arquitetura **U-Net** é composta por três componentes principais: o *encoder*, o gargalo (bottleneck) e o *decoder*. Esses três blocos estão conectados sequencialmente.

Além disso, as **U-Nets** também possuem o que são chamadas de *skip connections*. Elas conectam as entradas das funções de *Max Pooling* do *encoder* com as entradas do *decoder*, criando um fluxo direto de informação entre partes não sequenciais da rede.

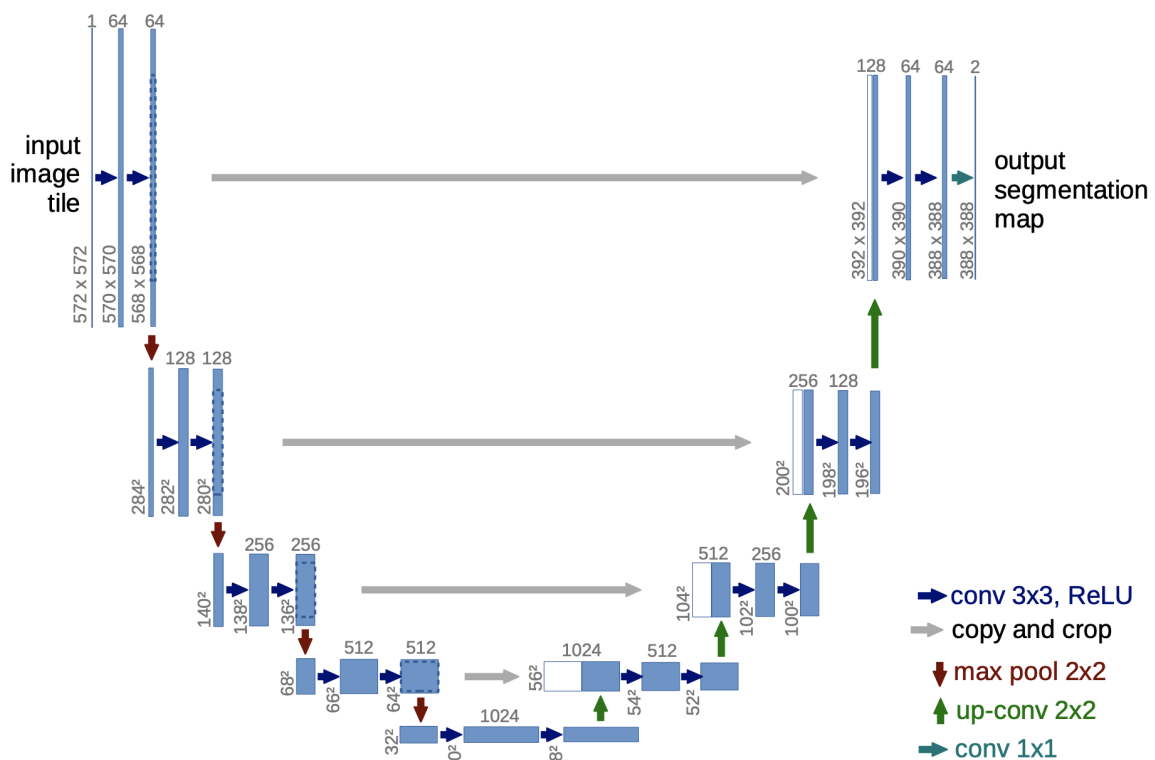


Figure 1: Imagem extraída do artigo original: "U-Net: Convolutional Networks for Biomedical Image Segmentation"[3]

Na figura 1, as setas cinzas denominadas *copy and crop* são as chamadas *skip connections*. Neste trabalho, cada sequência de convoluções 3x3 e *max pool 2x2* é denominada de **bloco de codificação** (*encoder block*).

Cada sequência de convoluções 3x3 e convolução transposta (*up-conv 2x2*) é denominada de **bloco de decodificação** (*decoder block*).

## Bloco de Codificação

O Bloco de Codificação da **U-Net** é responsável pela subamostragem da entrada (*downsampling*). Ele reduz a resolução da entrada (altura e largura) usando *max pooling* e aumenta seus canais utilizando convolução.

$$Conv1(x_0) = x_1 \rightarrow ReLU(x_1) = x_2 \rightarrow Conv2(x_2) = x_3 \rightarrow ReLU(x_3) = x_4 \rightarrow MaxPool(x_4) = x_5$$

## Gargalo (Bottleneck)

O gargalo da **U-Net** é responsável pela representação compacta da entrada. Ele mantém a resolução da entrada e aumenta seus canais utilizando convolução.

$$Conv1(x_0) = x_1 \rightarrow ReLU(x_1) = x_2 \rightarrow Conv2(x_2) = x_3 \rightarrow ReLU(x_3) = x_4$$

## Bloco de Decodificação

O Bloco de Decodificação da **U-Net** é responsável pela superamostragem da entrada (*upsampling*). Ele aumenta as dimensões da entrada (altura e largura) usando convolução transposta, ao mesmo tempo que reduz seus canais utilizando convolução.

$$TransposeConv1(x_0) = x_1 \rightarrow Conv1(x_1) = x_2 \rightarrow ReLU(x_2) = x_3 \rightarrow Conv2(x_3) = x_4 \rightarrow ReLU(x_4) = x_5$$

## 3.2 Estudo de Ablação sobre *Skip Connections* e Profundidade da Rede

A sequência de codificação, compactação e decodificação da arquitetura **U-Net** pode ser definida como abaixo de acordo com as definições apresentadas na última seção.

$$x_0 \rightarrow Encoder(x_0) = x_1 + skip_1 \rightarrow Encoder(x_1) = x_2 + skip_2 \rightarrow \dots \rightarrow Encoder(x_{n-1}) = x_n + skip_n \quad (1)$$

$$\rightarrow BottleNeck(x_n) = y_0 \quad (2)$$

$$\rightarrow Decoder(y_0 + skip_n s_n) = y_1 \rightarrow Decoder(y_1 + skip_{n-1} s_{n-1}) = y_2 \rightarrow \dots \rightarrow Decoder(y_{n-1} + skip_1 s_1) = y_n \quad (3)$$

Ou seja, a **U-Net** é uma sequência de  $n$  blocos de codificação, seguida de um único bloco de gargalo seguido por uma sequência de  $n$  blocos de decodificação.

Para investigar sistematicamente o impacto das *skip connections* e da profundidade da **U-Net** na performance de segmentação, realizamos um estudo de ablação abrangente utilizando a arquitetura **U-Net** proposta. Seja  $n$  o número de camadas de codificação/decodificação na rede, e seja  $s_i$  uma função indicadora binária definida para cada bloco do codificador  $i$ :

$$s_i = \begin{cases} 1 & \text{se a skip connection do bloco de codificação } i \text{ for utilizada} \\ 0 & \text{caso contrário} \end{cases}$$

Para cada profundidade  $n \in \{1, 2, 3, 4\}$ , todas as possíveis configurações de conexões skip foram avaliadas, resultando em  $2^n$  variantes distintas da rede. Cada variante corresponde a uma combinação única de conexões skip ativas e inativas, permitindo isolar a contribuição de cada caminho skip para a performance geral de segmentação.

Durante o treinamento, todas as variantes da rede foram otimizadas utilizando o mesmo conjunto de dados e hiperparâmetros, e métricas de avaliação, como coeficiente Dice e acurácia no conjunto de testes, foram registradas após cada época. O melhor modelo para cada configuração foi salvo, e os resultados foram agregados para analisar a influência tanto da profundidade  $n$  quanto do padrão de conexões skip  $s = (s_1, s_2, \dots, s_n)$  na qualidade da segmentação.

Esta metodologia permite uma avaliação detalhada dos componentes arquiteturais, fornecendo *insights* sobre a importância relativa das *skip connections* em diferentes camadas e os *trade-offs* associados à profundidade da rede.

Todas as redes foram treinadas em dois conjunto de dados: *DRIVE Digital Retinal Images for Vessel Extraction*[1] e *Skin Lesion Analysis Towards Melanoma Detection*[4].

## 4 Resultados

Nos resultados apresentados a seguir, a coluna *Skip Config* indica, de forma binária, quais *skip connections* foram utilizadas em cada bloco do codificador.

Cada caractere da *string* corresponde a um bloco, da camada mais próxima da entrada até a camada mais profunda, seguindo a definição:

$$s_i = \begin{cases} 1 & \text{se a skip connection do bloco } i \text{ foi utilizada} \\ 0 & \text{se a skip connection do bloco } i \text{ foi omitida} \end{cases}$$

Por exemplo, um resultado obtido para uma **U-Net** de profundidade  $n = 4$  (quatro blocos no codificador e decodificador), com apenas a primeira e peúltima *skip connection* ativas encontrase com *skip config* de 1010.

- $s_1 = 1 \rightarrow$  *skip connection* do primeiro bloco utilizada
- $s_2 = 0 \rightarrow$  *skip connection* do segundo bloco omitida
- $s_3 = 1 \rightarrow$  *skip connection* do terceiro bloco utilizada
- $s_4 = 0 \rightarrow$  *skip connection* do quarto bloco omitida

A melhor configuração de cada profundidade é destacada com fundo cinza.

#### 4.1 DRIVE Digital Retinal Images for Vessel Extraction

Depth ( $n$ )	Skip Config ( $s_1 \dots s_n$ )	Loss	Dice	Acc. on Test Set	Train Time (s)
1	0	0.0172	0.9920	0.9922	14.5
1	1	0.0153	0.9920	0.9919	12.3
2	00	0.0263	0.9876	0.9865	14.1
2	01	0.0173	0.9901	0.9894	15.8
2	10	0.0223	0.9883	0.9867	15.1
2	11	0.0148	0.9922	0.9919	15.7
3	000	0.0234	0.9896	0.9897	17.6
3	001	0.0109	0.9944	0.9947	18.2
3	010	0.0171	0.9917	0.9912	17.9
3	011	0.0136	0.9939	0.9949	18.8
3	100	0.0244	0.9870	0.9852	18.1
3	101	0.0123	0.9942	0.9949	18.7
3	110	0.0173	0.9912	0.9903	18.7
3	111	0.0136	0.9929	0.9933	19.2
4	0000	0.0383	0.9849	0.9872	24.1
4	0001	0.0128	0.9931	0.9933	24.8
4	0010	0.0163	0.9915	0.9905	25.1
4	0011	0.0163	0.9918	0.9916	24.4
4	0100	0.0290	0.9821	0.9769	24.4
4	0101	0.0119	0.9938	0.9943	25.4
4	0110	0.0162	0.9920	0.9919	25.1
4	0111	0.0126	0.9933	0.9932	28.0
4	1000	0.0200	0.9919	0.9926	26.6
4	1001	0.0124	0.9947*	0.9953	25.6
4	1010	0.0154	0.9905	0.9887	24.2
4	1011	0.0136	0.9928	0.9928	25.0
4	1100	0.0274	0.9849	0.9823	24.0
4	1101	0.0110	0.9938	0.9939	25.8
4	1110	0.0172	0.9912	0.9902	25.2
4 (U-Net Original)4	1111	0.0156	0.9919	0.9919	26.0

Table 1: Resultados do estudo de ablação: métricas na última época para cada combinação de profundidade *skip connections* para o *DRIVE Digital Retinal Images for Vessel Extraction*[1].

O melhor resultado de acurácia no conjunto de testes foi obtido na U-Net de profundidade  $n = 4$  (quatro blocos no codificador e decodificador), com apenas a primeira e última *skip connection* ativas (1001).

- $s_1 = 1 \rightarrow$  skip connection do primeiro bloco utilizada
- $s_2 = 0 \rightarrow$  skip connection do segundo bloco omitida
- $s_3 = 0 \rightarrow$  skip connection do terceiro bloco omitida
- $s_4 = 1 \rightarrow$  skip connection do quarto bloco utilizada

Observa-se que no caso do conjunto de dados [1], o uso de *skip connections* em todas as camadas de decodificação não resulta na melhor acurácia para nenhuma profundidade além da  $n = 2$ .

Curiosamente, a U-Net mais profunda ( $n = 4$ ) mas sem nenhuma *skip connection* (0000) obteve o pior resultado, indicando a importancia dessas conexões mesmo em redes mais profundas. Por outro lado, a rede mais profunda com todas as *skip connections* (1111) obteve resultado pior do que a rede mais rasa ( $n = 1$ ) sem nenhuma *skip connection*, indicando que existe um *trade-off* entre a profundidade da rede e a quantidade de *skip connections*.

## 4.2 Skin Lesion Analysis Towards Melanoma Detection

Depth ( $n$ )	Skip Config ( $s_1 \dots s_n$ )	Loss	Dice	Acc. on Test Set	Train Time (s)
1	0	0.3007	0.5727	0.8051	1012.9
1	1	0.3086	0.5770	0.8081	1045.9
2	00	0.2316	0.6392	0.8640	1124.9
2	01	0.2289	0.6361	0.8518	1155.5
2	10	0.2295	0.6280	0.8712	1146.6
2	11	0.2228	0.6145	0.8660	1174.8
3	000	0.2014	0.6607	0.8942	1237.1
3	001	0.2261	0.6454	0.8670	1235.4
3	010	0.1853	0.6619	0.9111	1220.8
3	011	0.1923	0.6431	0.9065	1221.8
3	100	0.1948	0.6638	0.9034	1206.7
3	101	0.1988	0.6571	0.8841	1276.8
3	110	0.1816	0.6578	0.9189	1253.9
3	111	0.2068	0.6616	0.8836	1286.6
4	0000	0.1634	0.6838	0.9251*	1399.0
4	0001	0.3304	0.5291	0.8061	1444.2
4	0010	0.1598	0.6588	0.9220	1393.3
4	0011	0.1842	0.6805	0.9008	1421.3
4	0100	0.1687	0.6650	0.9214	1396.8
4	0101	0.3395	0.5378	0.8296	1002.1
4	0110	0.1590	0.5838	0.9185	1426.3
4	0111	0.1607	0.6893	0.9141	1454.6
4	1000	0.1658	0.6893	0.9199	1404.7
4	1001	0.1820	0.6908	0.9063	1437.5
4	1010	0.1507	0.6977	0.9204	1486.5
4	1011	0.2788	0.6259	0.8468	1481.0
4	1100	0.1714	0.6419	0.9052	1485.2
4	1101	0.1639	0.7018	0.9201	1517.4
4	1110	0.1558	0.6403	0.9176	1494.1
4	1111	0.2673	0.6075	0.8708	1425.7

Table 2: Resultados do estudo de ablação: métricas na última época para cada combinação de profundidade *skip connections* para o *Skin Lesion Analysis Towards Melanoma Detection*[4].

O melhor resultado de acurácia no conjunto de testes foi obtido na U-Net de profundidade  $n = 4$  (quatro blocos no codificador e decodificador), com nenhuma *skip connection* ativas (0000).

- $s_1 = 0 \rightarrow$  skip connection do primeiro bloco omitida
- $s_2 = 0 \rightarrow$  skip connection do segundo bloco omitida
- $s_3 = 0 \rightarrow$  skip connection do terceiro bloco omitida
- $s_4 = 0 \rightarrow$  skip connection do quarto bloco omitida

Esse resultado demonstra uma rede profunda sem nenhuma *skip connection* superando todas as outras com ou sem tais conexões em todas configurações, inclusive a rede **U-Net** original.

## 5 Conclusão

Os resultados para ambos conjuntos de dados estudados demonstram que sim, há um *trade-off* entre as *skip connections* e o número de camadas de codificação/decodificação nas **U-Nets**. Em ambos os casos, o aumento da profundidade e *skip connections* não resultou em melhores resultados no conjunto de testes de maneira linear, demonstrando que o *trade-off* entre esses componentes é complexo.

Nos dois conjuntos de dados a profundidade  $n = 4$  foi a que trouxe melhor acurácia no conjunto de testes, demonstrando como a profundidade é diretamente relacionada com a qualidade da rede.

## References

- [1] LARXEL. Drive digital retinal images for vessel extraction, 2019. Kaggle, acessado em 19 out. 2025.
- [2] Vishal Rajput. Robustness of different loss functions and their impact on networks learning capability. *CoRR*, abs/2110.08322, 2021.
- [3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.
- [4] ZENITSU157. Skin lesion analysis towards melanoma detection, 2023. Kaggle, acessado em 19 out. 2025.