CentraleSupélec

U-Net: Convolutional Networks for Image Segmentation

AUTHORS

Marius Dragic
marius.dragic@student-cs.fr

February 28th 2025

1 Introduction

U-Net has revolutionised image segmentation with a design that is simple yet extremely efficient. Contrary to conventional approaches with massive amounts of datasets and enormous computational complexity, U-Net makes every example count with the multiple skip connections and intelligent exploitation of symmetry within the network. The design is efficient even with low amounts of datasets, an asset in the medical field where obtaining a high quantity of labeled pictures is not possible.

For this study, we chose to test the robustness of the model on the STARE dataset, a set of images designed for the segmentation of retinal blood vessels. This choice enables us to remain within a medical domain, while moving away from the initial context of U-Net, which had been designed for the segmentation of microscopic cellular structures. Segmenting retinal vessels poses specific challenges: these structures are thin, tangled and particularly numerous, which will test the model's ability to detect objects with complex, branching contours. We'll be assessing the model's ability to generalize to this new type of image, as the authors point out: "*We are sure that the U-Net architecture can be applied easily to many more tasks*" [1]. We will thus evaluate if U-Net model maintains its accuracy despite significant differences in textures and structures, illustrating the versatility of this architecture in a wide range of biomedical applications.

2 How U-Net works

2.1 U-Net architecture

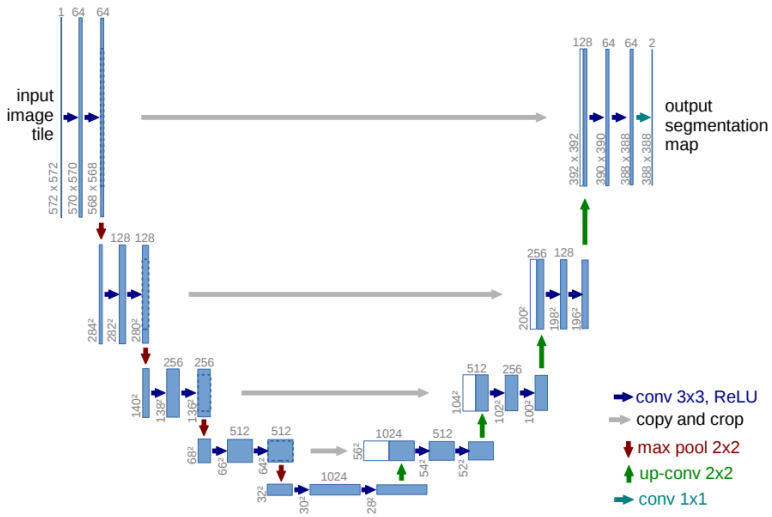


Figure 1: U-Net architecture

The U-Net structure consists of two main parts: a contracting block (or encoder) and an expanding block (or decoder), arranged symmetrically. The contracting block applies several 3×3 convolutions and 2×2 pooling operations to progressively reduce the spatial size of the image, while increasing the number of channels, thus facilitating the extraction of increasingly global features. The expansive block, for its part, uses up-convolutions

to restore lost resolution, chaining convolution and de-pooling to reconstruct details. U-Net's originality, however, lies in the use of skip connections, which directly link the stages of the contracting block to the corresponding stages of the expanding block. In practical terms, this enables precise local information (detected in the encoder) to be transferred to the decoder layers, thus avoiding the loss of fine detail that can occur during compression. At the end of this architecture, a 1×1 output layer performs the per-pixel classification.

2.2 Loss adapted to the segmentation of thin structures

Successful segmentation relies on a loss function adapted to the characteristics of the objects to be detected. In the case of the STARE dataset, the main difficulty lies in the extraction of blood vessels, which are fine and branched. Nevertheless a simple Binary Cross Entropy (BCE) loss can lead to the narrowest structures being poorly taken into account, often under-segmented or ignored by the model.

To overcome this limitation, we have introduced a Skeleton Recall Loss [2] in addition to the BCE weighted loss presented in the original paper [1], which forces the network to pay particular attention to tubular structures. This new loss is defined as follows, where Y_{skel} represents the skeletonized mask as shown in appendix 5.2:

Segmentation loss

$$\mathcal{L}_{\text{total}} = \underbrace{1 - \frac{1}{|C|} \sum_{c \in C} \frac{\sum_i Y_{\text{skel},i,c} \cdot \hat{Y}_{i,c}}{\sum_i Y_{\text{skel},i,c}}}_{\text{Skeleton Recall Loss}} + \lambda_{\text{BCE}} \cdot \underbrace{\mathbb{E}\left[w(x) \cdot \text{BCE}(x)\right]}_{\text{Weighted BCE loss}} \tag{1}$$

Skeleton Recall Loss (in red) compares predictions directly to vessel skeletons 5.2, rather than to full masks, thus preventing very fine structures from disappearing during learning. The addition of weighted BCE thanks to weight map generation described in appendix 5.1 ensures consistent overall segmentation, by compensating for class imbalances.

Thanks to this approach, the model learns to detect the finest vessels more accurately and to preserve their continuity, an essential element for medical applications such as the detection of vascular anomalies.

3 Experiment and results

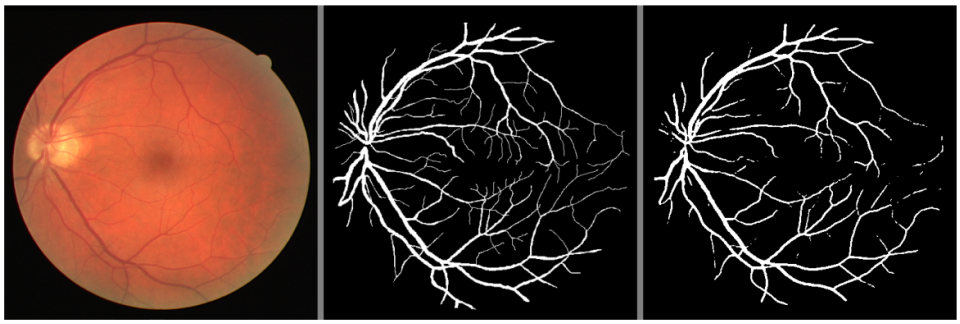
Model performance was evaluated on the STARE dataset, with particular attention paid to U-Net's ability to segment the retinal vascular network accurately. Two loss configurations were compared: on the one hand, the classic use of Binary Cross Entropy (BCE), and on the other hand, a combination of the weighted BCE loss from the original paper [1] and the skeleton recall loss from a research paper dating from 2024 [2], designed to favor the detection of tubular and filiform structures.

3.1 Segmentation results

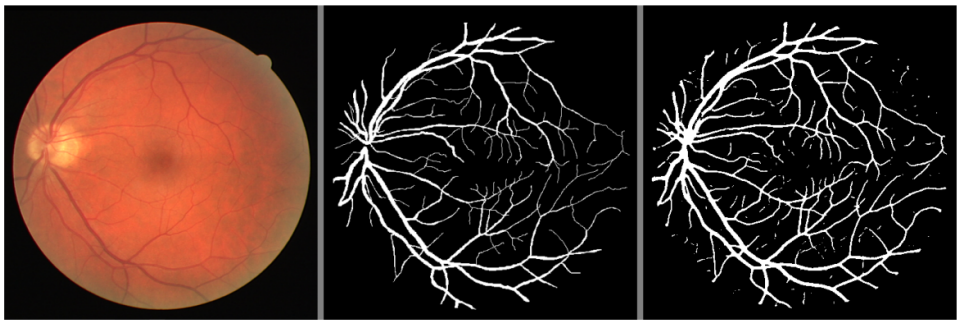
Looking at the segmentation results, its clear that using **BCE Loss** alone doesnt fully capture the vascular network. Indeed the main vessels are segmented correctly, but many

fine branches are either missing or broken, showing the model's struggle to detect the smallest details in retinal vessels. The final outpout mask seems subsegmented.

On the other hand, with **Skeleton Recall + Weighted BCE Loss**, the improvement is noticeable: both the thinnest and thickest vessels are better preserved, and their connectivity remains closer to the original structure. The new designed loss proposed here allow recognizing the thread-like nature of the vascular network and avoids the fragmentation issues seen with BCE Loss alone. This advantage is especially clear in areas with extensive branching, leading to a more anatomically consistent segmentation. However, some false positives appear along the edges of the image, slightly cluttering the final result, which would need post-processing to clean up.



(a) Segmentation with BCE Loss alone



(b) Segmentation with Skeleton Recall Loss + Weighted BCE Loss

Figure 2: Segmentation loss comparison

3.2 Performances evaluation

Performance evaluation confirms the value of our approach: by promoting better recall, we minimize the risk of under-segmentation of fine vessels, a critical issue in ophthalmology. In the medical field, it's preferable to identify as many structures as possible, even if it means tolerating a few false positives, rather than missing out on essential diagnostic elements. Skeleton Recall + Weighted BCE Loss thus preserves the continuity of the vascular network and detects details often overlooked by a conventional approach. This strategic choice guarantees more reliable segmentation that can be used in clinical practice, despite a loss of precision metric.

Metric	Skeleton + Weighted BCE	BCE only
Jaccard (%)	64.35	63.43
F1-score (%)	78.27	77.48
Recall (%)	91.12	74.54
Precision (%)	68.92	82.14
Accuracy (%)	95.59	96.28

Table 1: Performance comparison between Skeleton Recall Loss + Weighted BCE Loss and BCE alone.

4 Conclusion

This paper focuses on U-Net's superior performance in medical image segmentation, and in particular where detecting fine boundaries is crucial. Because U-Net is symmetrically configured and uses skip connections, it is able to catch not only global shape but also fine detail, a major asset in ophthalmology and other imaging. However, the segmentation of filiform structures poses specific challenges, requiring the adaptation of loss functions. The combination of skeleton recall with weighted BCE loss shows that targeted loss can significantly improve results, which is essential to avoid under-segmentation. More generally, these results confirm that segmentation success depends not only on the model and architecture used, but also on the loss design defined specifically for a given problem.

References

[1] Philipp Fischer Olaf Ronneberger and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (2015). URL: https://arxiv.org/abs/1505.04597.

[2] Yannick Kirchhoff et al. "Skeleton Recall Loss for Connectivity Conserving and Resource Efficient Segmentation of Thin Tubular Structures". In: (Dec. 2024). arXiv: 2404.03010 [cs.CV]. URL: https://arxiv.org/abs/2404.03010.

5 Appendix

5.1 Appendix A: Weight map

To target the most difficult areas to segment, U-Net original paper [1] introduces a weight per pixel, $w(x)$, into the cost function:

$$w(x) = w_c(x) + w_0 \exp\left(-\frac{(d_1(x) + d_2(x))^2}{2\sigma^2}\right) \tag{2}$$

where $l(x)$is the actual class of the pixel x. This weighting is designed to give greater importance to less-represented borders and regions. It is calculated using:

$$E = \sum_{x \in \Omega} w(x) \log\left(p_{\ell(x)}(x)\right) \tag{3}$$

where $w_c(x)$ corrects the imbalance between different classes, while $d_1(x)$ and $d_2(x)$ represent the distances from x to the regions of interest. In this way, pixels in contact with several vessels inherit a high weight, which encourages the network to better distinguish contours. The σ hyperparameter controls the spread of the weighting and $w0$ the maximum amplitude of the weighting.

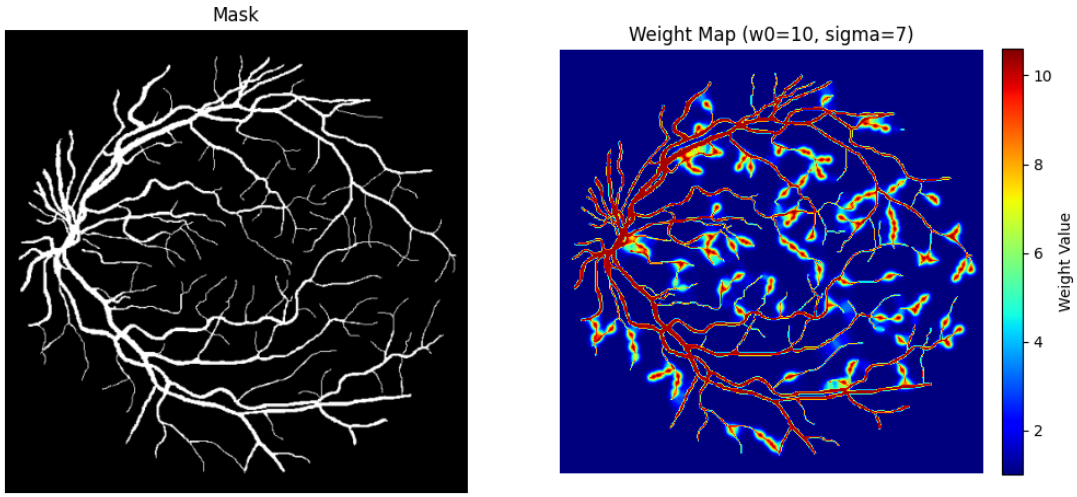


Figure 3: Mask Weight Map

5.2 Appendix B: Skeleton mask

The morphological operation of skeletonization reduces a binary structure to a thread-like, centered version, preserving its topology while removing redundant pixels. Applied to segmentation masks, it is particularly useful for analyzing fine structures such as blood vessels. In our case, Skeleton Recall Loss uses this transformation to compare the model prediction $\hat{Y}_{pred}$ with the skeletonized version of the actual mask Y_{skel}. This ensures that thin structures, often under-segmented by conventional models, are better taken into account, thus improving the recall metric.

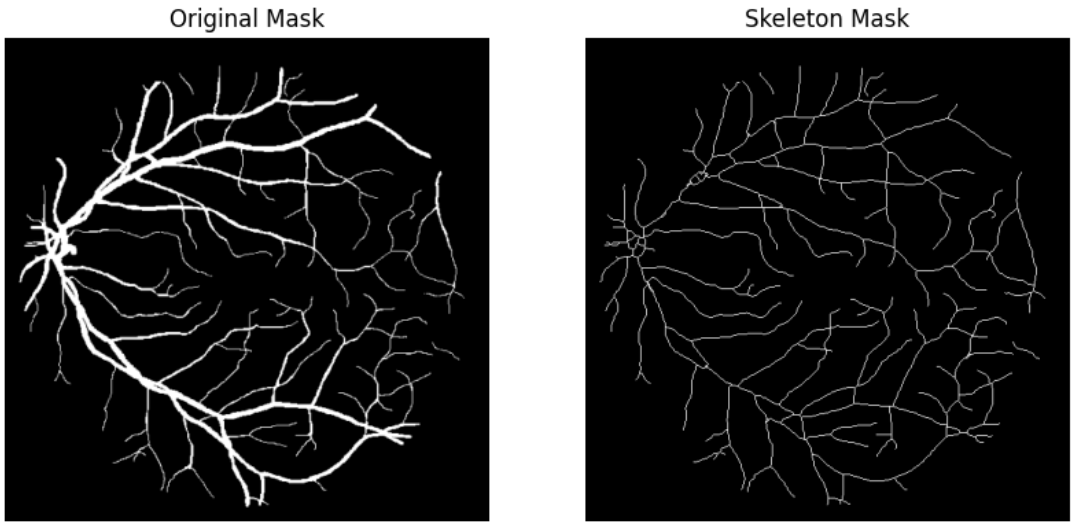


Figure 4: Skeleton Mask visualisation

5.3 Appendix C: Data augmentation

Data augmentation is an essential step in improving the robustness and generalization of deep learning models, particularly when a limited number of annotated images are available. Following the recommendations of section 3.1 of the original paper [1], we applied several random transformations to both the original image, the segmentation mask and the weight map. These transformations include horizontal and vertical flips, rotations, as well as resizing and slight deformations. The aim is to make the model more invariant to spatial variations, avoid overfitting and improve the network's ability to recognize vascular structures under different orientations and scales.

5.4 Appendix D: Hyperparameters

The model was trained on the dataset **STARE** initially composed of 80 train images of 512x512 pixels and 20 test images. With 3 augmentations per image, the dataset reaches 240 train images, enabling a considerable improvement in dataset size. Learning was performed on 25 epochs, using the Adam optimizer with an initial learning rate of 1e-4 and a reduction of this rate in the event of stagnation. Data were loaded in batches of 2 images, with binary segmentation optimized using our Skeleton Recall loss. Finally, a backup mechanism was implemented to retain the model offering the best performance over the validation set.

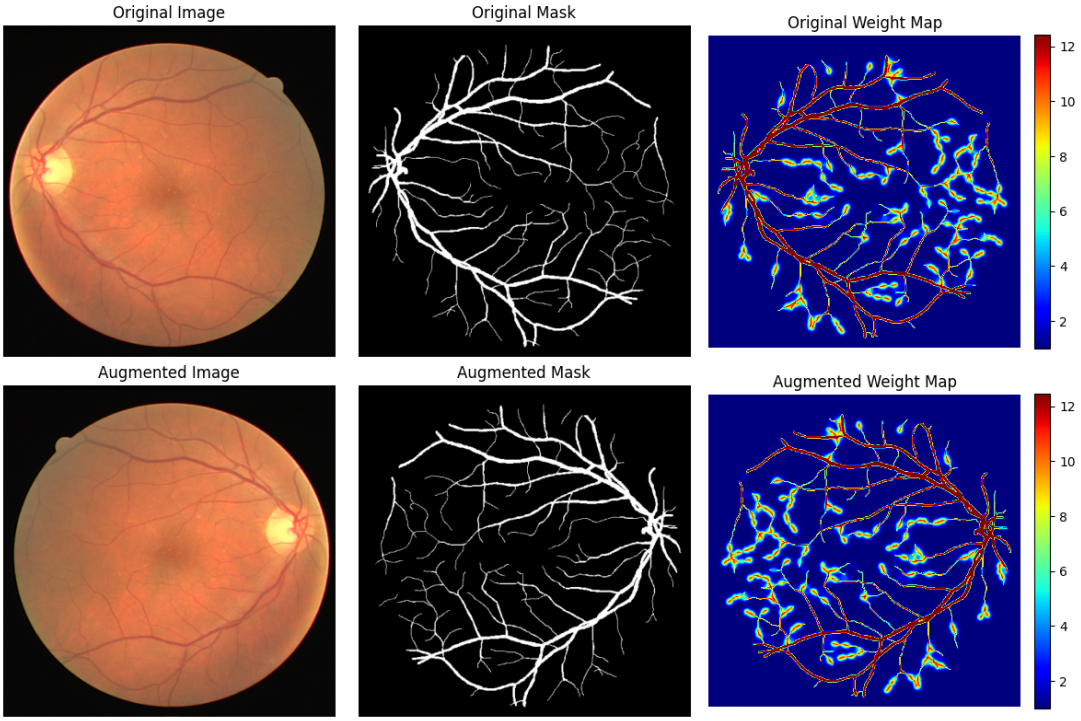


Figure 5: Augmented dataset

Hyperparameter	**Value**
Image size	512×512
Batch size	2
Optimizer	Adam
Initial learning rate	1×10^{-4}
Learning rate reduction	ReduceLROnPlateau (patience=5)
Number of epochs	25
Loss function	Skeleton Recall + Weighted BCE Loss
λ_{BCE}	5.0
σ Weight map	7.0
w_0 Weight map	10.0
Data augmentation	Rotations, Horizontal and Vertical flips
Checkpoint strategy	Best model on validation

Table 2: Summary of the hyperparameters used for training the model.

5.5 Appendix E: Skeleton Loss evolution

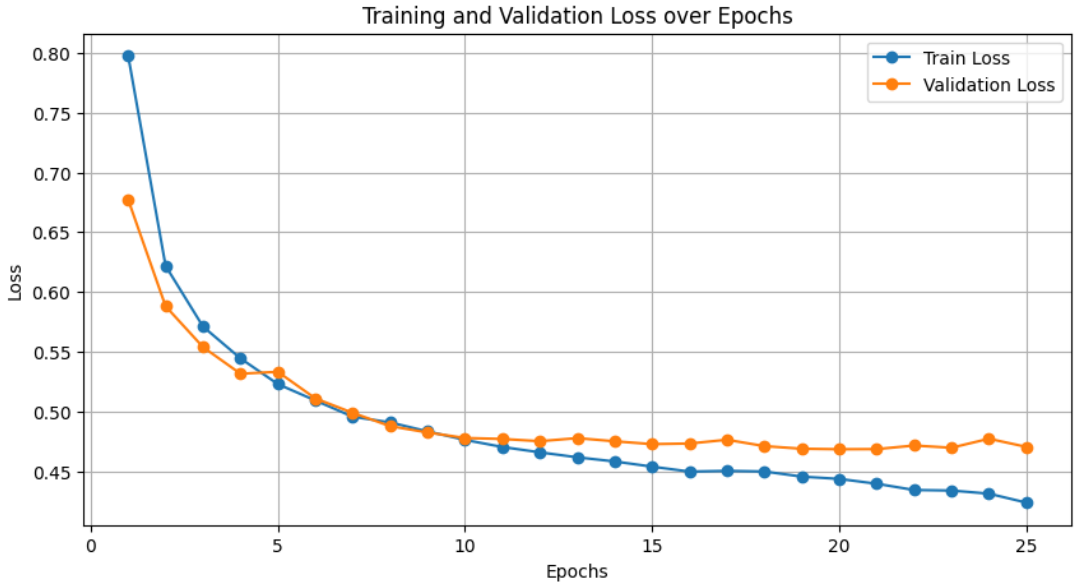


Figure 6: Augmented dataset

Validation and train loss seem to converge well after 25 epochs. A backup of the model associated with the best minimized loss on the test set has been implemented to avoid overfitting.