

# Lightweight Sclera Vessel Segmentation Using the IPAD Method

Matej Vitek<sup>1</sup>, Matevž Vidovič<sup>1</sup>, Žiga Emeršič<sup>1</sup>, Blaž Meden<sup>1</sup>,  
Peter Peer<sup>1</sup>, Kristijan Lenac<sup>2</sup>, Sandi Ljubić<sup>2</sup>

<sup>1</sup>Faculty of Computer and Information Science, University of Ljubljana,  
Večna pot 113, 1000 Ljubljana, Slovenia

<sup>2</sup>Faculty of Engineering, University of Rijeka,  
Vukovarska ulica 58, 51000 Rijeka, Croatia

E-mail: {matej.vitek, ziga.emersic, blaz.meden, peter.peer}@fri.uni-lj.si,  
mv9041@student.uni-lj.si, {kristijan.lenac, sandi.ljubic}@riteh.uniri.hr

## Abstract

Convolutional neural networks (CNNs) are typically overparameterised, presenting opportunities for model compression through pruning. We develop a flexible framework for iterative convolutional kernel pruning and apply it to U-Net and SegNet architectures trained for ocular vessel (vein) segmentation on the SBVPI dataset. Because vessel segmentation is complex and bounded by limited data, we employ a two-stage cascade using a lightweight sclera segmentation model for Region of Interest (ROI) extraction. Several kernel importance measures are evaluated, including  $L_1$  and  $L_2$  weight norms and the recently introduced Iterative Pruning with Activation Deviation (IPAD), as well as their affine combinations. Our results demonstrate that pruned vessel segmentation models substantially outperform compact models trained from scratch, achieving improvements of up to 10 percentage points in IoU. Furthermore, while activation-based pruning generally exhibits reduced consistency under severe data constraints, the IPAD method remains highly competitive for vessels. In specific configurations, such as with SegNet, IPAD<sub>2</sub> successfully outperforms traditional  $L_2$  weight norms, demonstrating its viability even when validation data is extremely limited. Even at aggressive pruning levels – retaining only 3% of original FLOPs – competitive segmentation performance is maintained.

## 1 Introduction

Convolutional neural networks [1] have become the cornerstone of modern visual recognition systems, owing to their ability to automatically learn hierarchical feature representations. Despite their effectiveness, CNNs suffer from substantial computational demands in terms of complexity, memory usage, and energy consumption – limitations that pose significant challenges in edge deployment scenarios such as biometric recognition on mobile devices [2, 3].

Interestingly, many deep CNNs are heavily overparameterised relative to the complexity of their target tasks, as it has been shown that a small fraction of a trained network’s weights can be used to predict the remaining ones [4]. This redundancy motivates model compression techniques [5], among which pruning [6, 7] has emerged as a particularly effective approach. Pruning identifies and removes weights or convolutional kernels that contribute least to

the network’s output, reducing computational complexity while potentially even improving performance through reduced overfitting [8].

Common kernel importance criteria include the  $L_1$  [9] and  $L_2$  [10] norms of kernel weights, which assume that smaller weights contribute less to the output. However, comparing weight magnitudes across layers with differing roles and input distributions is inherently unreliable. The Iterative Pruning with Activation Deviation (IPAD) method [8] addresses this limitation by assessing kernel importance based on the deviation of its activations from the layer mean, operating closer to the kernels’ actual influence on the network’s behaviour.

In this work, we develop a flexible framework for iterative kernel pruning and apply it to U-Net [11] and SegNet [12] architectures designed for high-resolution ocular vessel segmentation using the SBVPI dataset [13]. Because vessel segmentation is highly complex and bounded by limited annotation data, we employ a two-stage cascade using a lightweight sclera segmentation model for Region of Interest (ROI) extraction. We evaluate affine combinations of weight-based and IPAD importance measures, investigating how the IPAD criterion performs under data-constrained conditions compared to uniform and random pruning baselines. The remainder of this paper is organised as follows: Section 2 describes the pruning pipeline, importance measures, and experimental setup; Section 3 presents the results; Section 4 discusses the implications; and Section 5 concludes the paper.

## 2 Methodology

### 2.1 Iterative Pruning Pipeline

Our pruning pipeline follows a global iterative approach. Beginning with a fully trained model, the importance of every convolutional kernel in the network is computed using the chosen measure. The 100 least important kernels are then removed, and the model’s floating-point operations (FLOPs) are recalculated using

$$\text{FLOPs} = K_W \cdot K_H \cdot K_D \cdot N_K \cdot I_W \cdot I_H, \quad (1)$$

where  $K_W$ ,  $K_H$ , and  $K_D$  are the kernel width, height, and depth,  $N_K$  is the number of kernels, and  $I_W$ ,  $I_H$  are the input dimensions. This removal–recalculation cycle repeats until the cumulative FLOP reduction exceeds 1%

of the initial value, at which point the model is retrained. Pruning and retraining cycles continue until the desired FLOPs target is reached.

To prevent excessive pruning of individual layers – which would create feature-propagation bottlenecks – we impose the constraint  $F_{\text{layer}} > \min(20\%, 0.8 \cdot F_{\text{network}})$ , where  $F_{\text{layer}}$  and  $F_{\text{network}}$  denote the percentages of FLOPs retained by the layer and the entire network, respectively. This allows the minimum layer threshold to relax as the network becomes heavily pruned.

## 2.2 Kernel Importance Measures

We evaluate several kernel importance measures. The  $L_p$  norm of a kernel’s weight matrix  $\mathbf{W}$  serves as a simple proxy for importance:

$$\|\mathbf{W}\|_p = \left( \sum_{i=1}^n |w_i|^p \right)^{1/p}, \quad p \in \{1, 2\}. \quad (2)$$

The IPAD method [8] instead assesses importance through the Activation Deviation Criterion (ADC). For each kernel  $k$ , its average activation  $\bar{\mathbf{A}}_k$  is computed over the validation set. The average layer activation  $\bar{\mathbf{S}}_l$  is obtained by averaging  $\bar{\mathbf{A}}_k$  across all kernels in layer  $l$ . The ADC is then defined as:

$$\text{ADC}_{k,m} = \frac{\|\bar{\mathbf{A}}_k - \bar{\mathbf{S}}_l\|_m}{h \cdot w}, \quad (3)$$

where  $m$  is the choice of  $L$ -norm and  $h \times w$  are the spatial dimensions of the activation. Kernels whose activations deviate more from the layer mean are deemed more important, as they produce outputs that are less replaceable by other kernels within the same layer.

Weight-based and IPAD measures are combined via affine combinations:

$$I_k = \alpha \cdot \|\mathbf{W}_k\|_p + (1 - \alpha) \cdot \text{ADC}_{k,p}, \quad (4)$$

where  $p \in \{1, 2\}$ , and  $\alpha \in \{0, 0.1, 0.5, 0.9, 1\}$ . The pairing of  $L_p$  with  $\text{ADC}_p$  (using the same  $p$ ) follows prior work [8]. We denote by  $\text{IPAD}_p^*$  the combination that achieves the best IoU on the validation set over all tested  $\alpha$  values. Uniform and random pruning serve as baselines.

## 2.3 Experimental Setup

We employed U-Net [11] and SegNet [12] – two established encoder–decoder architectures for semantic segmentation. U-Net uses skip connections to propagate fine spatial detail from the encoder to the decoder, while SegNet transfers max-pooling indices instead. Both are well suited for high-resolution segmentation of small structures such as ocular veins. Table 1 summarises the architectural configurations.

The models were trained on the SBVPI dataset [13], which provides high-resolution ocular images with pixel-level annotations: 1,858 images for sclera and 128 images for veins (a subset of 23 individuals). A 70/20/10 train/validation/test split was used. All images were resized to uniform dimensions (Table 1) using Lanczos interpolation. Data augmentation (random affine transforms and colour jitter) was applied for vein segmentation only.

Table 1: Architectural parameters of the models used.

| Model         | $K_0$ | $f$ | $d$ | GFLOPs | Input              |
|---------------|-------|-----|-----|--------|--------------------|
| SegNet sclera | 16    | 2.0 | 4   | 40.61  | $1024^2$           |
| U-Net sclera  | 16    | 2.0 | 6   | 68.13  | $1024^2$           |
| SegNet veins  | 32    | 1.5 | 4   | 138.37 | $1024 \times 2048$ |
| U-Net veins   | 32    | 2.0 | 4   | 378.82 | $1024 \times 2048$ |

$K_0$  – initial kernels;  $f$  – expansion factor;  $d$  – depth.

For vein segmentation, we adopted a Region of Interest (ROI) extraction cascade inspired by prior work [2]: a lightweight U-Net ( $\sim 3$  GFLOPs) first segments the sclera and its output mask suppresses irrelevant non-scleral regions in the vein segmentation input. All models were trained using the Adam optimiser [14] with a learning rate of  $10^{-4}$  and weighted Multi-Class Dice Loss [15], for a fixed number of epochs without early stopping; the resulting overparameterisation of the base models is expected and is precisely the motivation for subsequent pruning. After each pruning cycle, the retained weights are preserved and used as initialisation for continued training (fine-tuning), directly leveraging the learned representations rather than restarting from scratch. The reduced model capacity after pruning acts as an implicit regulariser, limiting overfitting despite the extended training. For each importance measure, intermediate models at 75%, 50%, 25%, and 3% of original FLOPs were saved and subsequently trained to an equal total number of epochs to ensure fair comparison.

## 3 Results

### 3.1 IPAD Effectiveness

Table 2 presents IoU scores at 25% of original FLOPs for all model–task configurations. For sclera segmentation, the optimal  $\text{IPAD}_1^*$  combination outperforms the standalone  $L_1$  measure by 2.11 pp for SegNet and 0.26 pp for U-Net. In SegNet, the highest performances are obtained by either relying entirely on the pure IPAD criterion ( $\alpha = 0$ , yielding 90.35% for  $\text{IPAD}_1^*$ ) or focusing primarily on it ( $\alpha = 0.1$ , yielding 90.40% for  $\text{IPAD}_2^*$ ). This demonstrates that activation-based importance can surpass weight-based methods when sufficient validation data is available.

For vessel segmentation, the effectiveness of IPAD varies across architectures due to the substantially smaller validation set (26 images vs. 372 for sclera). While  $\text{IPAD}_1^*$  defaults to the standalone  $L_1$  norm ( $\alpha = 1.0$ ) for both U-Net and SegNet at 25% FLOPs (as shown in Table 2), other configurations tell a different story. Notably, the  $\text{IPAD}_2^*$  measure successfully outperformed the standalone  $L_2$  norm when applied to SegNet for vessel segmentation. Furthermore, across all pruning levels and network architectures, the IPAD-augmented measures outperformed standalone  $L$ -norms in a majority of evaluated scenarios for vessels (10 out of 16 cases). This demonstrates that activation-based importance can remain competitive and even surpass weight-based methods on highly complex

tasks, despite severely limited validation data. Finally, all pruning measures significantly outperform or match the unpruned baseline on the vessel task, confirming substantial overparameterisation. Although the absolute IoU scores for vessels peak at just over 40%, this is largely an artefact of the pixel-wise evaluation metric rather than a failure of the models to capture vessel topology (see Section 4 for a detailed discussion).

Table 2: IoU (%) at 25% of original FLOPs. Best pruned result per row in **bold**. “Small” denotes a compact model trained from scratch at  $\sim 25\%$  FLOPs.

|        |    | Unpr. | Small | $L_1$        | $IPAD_1^*$   | $L_2$ | $IPAD_2^*$   | Unif. | Rand. |
|--------|----|-------|-------|--------------|--------------|-------|--------------|-------|-------|
| Sclera | SN | 90.74 | 89.50 | 88.24        | 90.35        | 89.59 | <b>90.40</b> | 90.06 | 89.44 |
|        | UN | 90.64 | 88.80 | 90.68        | <b>90.94</b> | 90.50 | 90.61        | 88.63 | 88.94 |
| Veins  | SN | 37.23 | 37.75 | <b>37.89</b> | <b>37.89</b> | 36.35 | 37.12        | 33.58 | 36.81 |
|        | UN | 38.18 | 33.72 | <b>40.84</b> | <b>40.84</b> | 40.61 | 40.61        | 40.31 | 39.34 |

SN – SegNet; UN – U-Net; Unpr. – unpruned baseline.  $IPAD^*$  indicates the optimal  $\alpha$  combination.

### 3.2 Performance Across Pruning Levels

Figure 1 illustrates the relationship between the retained FLOPs and IoU for U-Net on both segmentation tasks. For sclera, all measures maintain performance comparable to the unpruned baseline down to 25% of the original FLOPs, with  $IPAD_1$  ( $\alpha = 0.1$ ) reaching the highest score at 75%. Performance degrades more substantially at 3% FLOPs, though IoU remains above 79%.

For vein segmentation, pruned models consistently outperform the unpruned baseline at moderate pruning levels (25–75% FLOPs), corroborating findings from the original IPAD study [8]. Performance at 3% FLOPs exhibits high variability across measures, with uniform pruning surprisingly outperforming all others. Interestingly,  $IPAD_1$  at  $\alpha = 0.1$  – which excels for sclera – performs among the poorest for veins, further underscoring its dependence on validation set size.

### 3.3 Impact of $\alpha$ Choice

We investigated how the selection of the hyperparameter  $\alpha$  influences pruning performance. As shown in Table 3, we observed a distinct V-shaped pattern across most tasks, indicating that extreme values of  $\alpha$  (close to 0 or 1) generally yield better results than intermediate values. This suggests that the affine combination may not be the most effective fusion strategy. For instance, in SegNet veins, performance drops from 37.59% at  $\alpha = 0$  to 35.54% at  $\alpha = 0.1$ , before rising again to 37.89% at  $\alpha = 1.0$ . Furthermore, using pure activation deviation ( $IPAD_2$  with  $\alpha = 0$ ) in U-Net for veins resulted in a complete loss of performance (IoU of 0%, not shown). This implies that relying solely on activation deviations can be highly unstable when the validation set is extremely small.

### 3.4 Segmentation with Smaller Models

Pruned models generally outperform compact models trained from scratch at the same FLOP level, validating the benefits of training a larger network and subsequently

Table 3: Effect of  $\alpha$  on IoU (%) using  $IPAD_1$  at 25% FLOPs.

|        |    | $\alpha = 0$ | $\alpha = 0.1$ | $\alpha = 0.5$ | $\alpha = 0.9$ | $\alpha = 1.0$ |
|--------|----|--------------|----------------|----------------|----------------|----------------|
| Sclera | SN | 90.35        | 90.20          | 89.55          | 87.01          | 88.24          |
|        | UN | 90.14        | 90.39          | 90.09          | 90.94          | 90.68          |
| Veins  | SN | 37.59        | 35.54          | 36.04          | 37.84          | 37.89          |
|        | UN | 39.87        | 38.34          | 40.26          | 39.83          | 40.84          |

pruning it. Across all evaluated 25% and 3% FLOP scenarios, the optimally pruned model surpassed its compact counterpart in 7 out of 8 cases (see Table 2 for the 25% subset). For U-Net veins at 25% FLOPs, the pruned model achieves 40.84% IoU compared to 33.72% for the compact model – a difference of over 7 pp. At 3% FLOPs, SegNet veins shows an even more pronounced advantage of 10 pp (33.34% vs. 23.34%).

For sclera segmentation, however, the advantage of pruning is less pronounced, yielding only marginal performance gains. Notably, the compact SegNet model configured to 3% of original FLOPs slightly outperformed the corresponding pruned model by 2.12%. This anomaly may be partially attributed to architectural constraints that resulted in the compact model having roughly 27% more actual FLOPs than the 3% target, as well as the inherent instability associated with pruning to extremely low FLOP levels where removing a single critically important kernel can significantly degrade performance. Overall, however, the results confirm that pruning preserves complex feature representations that small architectures are typically unable to learn independently.

### 3.5 Classical Segmentation Methods

To contextualise the performance of our deep CNN-based segmentation models, we compare them with two classical non-deep learning vessel segmentation approaches: B-COSFIRE and Coye filtering.

B-COSFIRE (Bar-selective Combination Of Shifted Filter Responses) [16] is an advanced image processing algorithm designed for the segmentation of elongated structures, such as blood vessels in retinal images. It utilises symmetric and asymmetric filters that are selective for bars and bar-endings. The Coye filtering algorithm [17] employs Principal Component Analysis (PCA) to convert RGB images to grayscale, followed by Contrast-Limited Adaptive Histogram Equalisation (CLAHE) to improve contrast. It then utilises Isodata thresholding to differentiate vessel structures, followed by morphological operations to refine the segmentation by removing small, non-vessel pixels.

Both classical segmentation methods performed significantly worse than all CNN-based approaches, even when compared to the most aggressively pruned models (retaining only 3% FLOPs), as shown in Table 4. B-COSFIRE achieved an IoU of 16.32%, while Coye filtering achieved only 10.34%. This stark contrast underscores the superiority of automatic hierarchical feature extraction over handcrafted feature engineering for the complex task of ocular vein segmentation. We hypothesise that the original application domains for these classical filters were

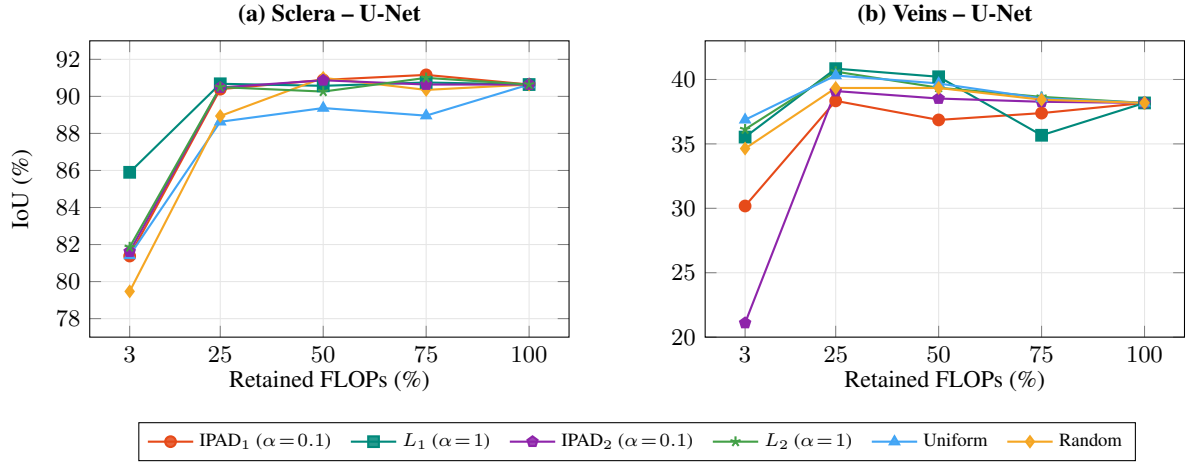


Figure 1: IoU scores for sclera (a) and vein (b) segmentation using U-Net at various levels of retained FLOPs. For the IPAD measures,  $\alpha = 0.1$  is used.

too dissimilar to the specific characteristics of the SBVPI dataset, contributing to their poor overall performance.

Table 4: Performance of classical vein segmentation methods on SBVPI.

| Method         | IoU (%) | F1 (%) |
|----------------|---------|--------|
| B-COSFIRE      | 16.32   | 28.06  |
| Coye filtering | 10.34   | 18.74  |

## 4 Discussion

Our results highlight the dual challenge of data scarcity in the vessel segmentation task: the limited dataset restricts generalisation, while the small validation set (26 images vs. 372 for sclera) yields noisier activation estimates for IPAD. Despite this, IPAD-augmented measures remained highly competitive and even outperformed standalone  $L_2$  norms in specific configurations, indicating that IPAD retains its viability under severe data constraints.

Regarding the  $\alpha$  hyperparameter, the V-shaped pattern observed in Table 3 suggests that the affine combination may not be the most effective fusion strategy. Alternative formulations using different  $L_k$  norms over the paired scores warrant investigation: a high  $k$  would prioritise pruning kernels where both criteria assign moderate importance, while a negative  $k$  would target kernels where either criterion assigns low importance. In practice,  $\alpha$  must be treated as a task-specific hyperparameter, selected on the basis of validation performance.

The seemingly low IoU values for vein segmentation (maximum 40.84%) are partly an artefact of the evaluation metric. As illustrated in Figure 2, predictions that correctly capture the vessel topology are penalised by pixel-wise metrics like IoU due to thickness discrepancies – a consequence of the gradual transition of vein boundaries into the sclera, which makes consistent thickness annotation inherently difficult. For biometric recognition, the overall vessel structure matters more than pixel-perfect boundary

adherence, underscoring the need to evaluate such models on downstream recognition performance.

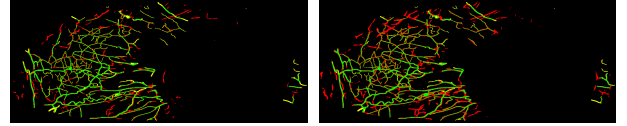


Figure 2: Vein segmentation colourmaps for a pruned  $L_1$  model (left) vs. a compact model trained from scratch (right). Green: TP; red: FP; yellow: FN. Despite capturing vessel topology, thickness discrepancies reduce IoU. Best viewed in colour and zoomed in.

## 5 Conclusion

We presented a flexible framework for iterative CNN kernel pruning and demonstrated its effectiveness on ocular sclera and vein segmentation tasks. Our key findings are: (1) pruned models substantially outperform compact models trained from scratch, especially for vein segmentation; (2) while the IPAD pruning criterion’s consistency depends on validation set size, it excels with sufficient data and remains a highly competitive alternative to traditional  $L$ -norms even under severe data constraints; and (3) competitive performance is maintained even at extreme pruning levels of 3% of the original FLOPs.

Future work will explore alternative methods for combining kernel importance measures (e.g. different  $L_k$  norm formulations), investigate higher-order variations of the ADC criterion, evaluate the impact of segmentation quality on downstream biometric recognition accuracy, and compare pruning with knowledge distillation [18] for model compression. Both our pruning framework and the code for the complete experiment are publicly available at [https://github.com/MatevzVidovic/CNN\\_kernel\\_pruning](https://github.com/MatevzVidovic/CNN_kernel_pruning) and <https://github.com/MatejVitek/ERK-2026-IPAD>.

## References

- [1] Y. Lecun et al. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.
- [2] Peter Rot et al. “Deep Sclera Segmentation and Recognition”. In: *Handbook of Vascular Biometrics (HVB)*. Ed. by Andreas Uhl et al. Springer, 2020, pp. 395–432.
- [3] Matej Vitek, Vitomir Štruc, and Peter Peer. “GazeNet: A lightweight multitask sclera feature extractor”. In: *Alexandria Engineering Journal* 112 (2025), pp. 661–671.
- [4] Misha Denil et al. “Predicting Parameters in Deep Learning”. In: *Neural Information Processing Systems*. 2013.
- [5] Tejalal Choudhary et al. “A comprehensive survey on model compression and acceleration”. In: *Artificial Intelligence Review* 53.7 (Oct. 2020), pp. 5113–5155. ISSN: 1573-7462.
- [6] Yann LeCun, John Denker, and Sara Solla. “Optimal Brain Damage”. In: *Advances in Neural Information Processing Systems*. Ed. by D. Touretzky. Vol. 2. Morgan-Kaufmann, 1989.
- [7] Tailin Liang et al. “Pruning and quantization for deep neural network acceleration: A survey”. In: *Neurocomputing* 461 (2021), pp. 370–403. ISSN: 0925-2312.
- [8] Matej Vitek et al. “IPAD: Iterative pruning with activation deviation for sclera biometrics”. In: *Journal of King Saud University - Computer and Information Sciences* 35.8 (2023), p. 101630. ISSN: 1319-1578.
- [9] Hao Li et al. “Pruning Filters for Efficient ConvNets”. In: *ArXiv abs/1608.08710* (2016).
- [10] Yang He et al. “Soft Filter Pruning for Accelerating Deep Convolutional Neural Networks”. In: *International Joint Conference on Artificial Intelligence*. 2018.
- [11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: *CoRR abs/1505.04597* (2015).
- [12] Vijay Badrinarayanan, Ankur Handa, and Roberto Cipolla. “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Robust Semantic Pixel-Wise Labelling”. In: *ArXiv abs/1505.07293* (2015).
- [13] Matej Vitek et al. “A Comprehensive Investigation into Sclera Biometrics: A Novel Dataset and Performance Study”. In: *Neural Computing & Applications (NCAA)* 32 (2020), pp. 17941–17955.
- [14] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2017.
- [15] Reza Azad et al. *Loss Functions in the Era of Semantic Segmentation: A Survey and Outlook*. 2023.
- [16] George Azzopardi et al. “Trainable COSFIRE filters for vessel delineation with application to retinal images”. In: *Medical Image Analysis* 19.1 (2015), pp. 46–57. ISSN: 1361-8415.
- [17] Tyler Coye. *A Novel Retinal Blood Vessel Segmentation Algorithm for Fundus Images*. MATLAB Central File Exchange; retrieved April 2, 2025. 2015.
- [18] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. *Distilling the Knowledge in a Neural Network*. 2015.