Journal of King Saud University – Computer and Information Sciences 35 (2023) 101630

HOSTED BY

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

IPAD: Iterative pruning with activation deviation for sclera biometrics

Matej Vitek [a,*], Matic Bizjak [a], Peter Peer [a], Vitomir Štruc [b]

[a] Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, SI-1000 Ljubljana, Slovenia
[b] Faculty of Electrical Engineering, University of Ljubljana, Tržaška 25, SI-1000 Ljubljana, Slovenia

ARTICLE INFO

Article history:
Received 23 January 2023
Revised 2 June 2023
Accepted 16 June 2023
Available online 28 June 2023

Keywords:
Biometrics
Sclera segmentation
Ocular biometrics
Ocular segmentation
Model pruning
Lightweight deep learning

ABSTRACT

The sclera has recently been gaining attention as a biometric modality due to its various desirable characteristics. A key step in any type of ocular biometric recognition, including sclera recognition, is the segmentation of the relevant part(s) of the eye. However, the high computational complexity of the (deep) segmentation models used in this task can limit their applicability on resource-constrained devices such as smartphones or head-mounted displays. As these devices are a common desired target for such biometric systems, lightweight solutions for ocular segmentation are critically needed. To address this issue, this paper introduces IPAD (Iterative Pruning with Activation Deviation), a novel method for developing lightweight convolutional networks, that is based on model pruning. IPAD uses a novel filter-activation-based criterion (ADC) to determine low-importance filters and employs an iterative model pruning procedure to derive the final lightweight model. To evaluate the proposed pruning procedure, we conduct extensive experiments with two diverse segmentation models, over four publicly available datasets (SBVPI, SLD, SMD and MOBIUS), in four distinct problem configurations and in comparison to state-of-the-art methods from the literature. The results of the experiments show that the proposed filter-importance criterion outperforms the standard L^1 and L^2 approaches from the literature. Furthermore, the results also suggest that: (*i*) the pruned models are able to retain (or even improve on) the performance of the unpruned originals, as long as they are not over-pruned, with RITnet and U-Net at 50% of their original FLOPs reaching up to 4% and 7% higher IoU values than their unpruned versions, respectively, (*ii*) smaller models require more careful pruning, as the pruning process can hurt the model's generalization capabilities, and (*iii*) the novel criterion most convincingly outperforms the classic approaches when sufficient training data is available, implying that the abundance of data leads to more robust activation-based importance computation.

1. Introduction

Sclera biometrics is a subfield of biometric identity recognition research. It studies the recognition of individuals using traits from the sclera vasculature, i.e., the vasculature contained in the white portion of the human eye (Vitek et al., 2020a; Das et al., 2013). Unlike competing ocular biometric modalities, such as the retina, the vasculature of the sclera is a visible ocular characteristic and, therefore, does not require specialized acquisition hardware for the imaging process, which makes it suitable for everyday applications. Furthermore, it can be imaged in the visible spectrum (VIS) and is not affected by the presence of eye lenses, unlike the iris (Derakhshani and Ross, 2007; Rot et al., 2018). These characteristics make sclera recognition an ideal candidate for mobile authentication schemes, either as a standalone modality or as part of multi-modal authentication solutions. However, much of the recent sclera-biometrics research has focused primarily on model accuracy and has largely ignored the memory footprint and computational complexity of the processing pipelines. This makes the results of such research difficult to apply to mobile (and edge) authentication schemes in practice, and provides strong motivation for the development of lightweight models and mechanisms capable of reducing the (time/space) complexity of the overall recognition pipelines. The development of such lightweight models is the main goal of the research work presented in this paper.

* Corresponding author.
E-mail address: matej.vitek@fri.uni-lj.si (M. Vitek).
URL: https://sclera.fri.uni-lj.si/ (M. Vitek).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2023.101630
1319-1578/

A vital part of the sclera recognition process, that also accounts for a significant part of the overall (computation and memory) load, is sclera segmentation. Recent research focusing on comparative performance evaluations of sclera segmentation models (Das et al., 2018; Das et al., 2019; Vitek et al., 2020b; Vitek et al., 2023) has demonstrated the superiority of deep learning solutions for this task. However, the top performing models based on general purpose architectures, such as U-Net (Ronneberger et al., 2015) and DeepLab (Chen et al., 2018), are typically over-parameterized and, as a result, are quite demanding with respect to the hardware needed for processing, both in terms of memory footprint as well as the number of operations required for real-time processing. To address some of these challenges, the OpenEDS competition was organized recently (Garbin et al., 2019). The goal of the competition was to design ocular segmentation models that could be run on modest hardware (available with virtual-reality (VR) head mounted displays) and to encourage research into lightweight model-design strategies. While several participants entered the competition, the most successful models included hand-designed architectures (that met the 1 MB memory-footprint constraint of OpenEDS) following the common, U-Net inspired, encoder-decoder network topology (Chaudhary et al., 2019; Perry and Fernandez, 2019; Boutros et al., 2019). The best performing of these models, RITnet (Chaudhary et al., 2019), featured only around 250000 parameters, but due to the hand-crafted design still led to a suboptimal trade-off between model complexity and performance – as also demonstrated in the experimental part of this paper.

A more structured approach towards reducing the memory footprint and computational complexity on contemporary deep learning models is to adopt solutions from the field of model compression (Choudhary et al., 2020). While different techniques have been proposed in the literature in this area over the years, including knowledge distillation procedures (Hinton et al., 2014; Romero et al., 2015; Zhang et al., 2018; Schmid et al., 2023), quantization mechanisms (Zhou et al., 2017; Zeng et al., 2022; Nevarez et al., 2023), low-rank approximation techniques (Chang et al., 2022; Kozyrskiy and Phan, 2020), weight-sharing strategies (Yi et al., 2017; Dupuis et al., 2021; Dupuis et al., 2022) and approximate-computation schemes (Kim et al., 2018; Masadeh et al., 2018; Hu et al., 2022), one of the most popular and generally applicable solutions towards developing lightweight models for various computer-vision tasks, that also alleviates the need for (sub-optimal) hand-crafted model architectures and is also at the core of this work, is *model pruning* (LeCun et al., 1990; Liang et al., 2021). Model pruning starts with a large model and removes low-impact filters (or neurons) to decrease its complexity while keeping the accuracy as high as possible. However, the mechanisms that decide which filters of a convolutional neural network to prune remain underexplored, with most of the existing approaches relying on simplistic L^1 (Li et al., 2017) or L^2 (He et al., 2018; Chin et al., 2020) norms of the kernel weights. As noted by He et al. (2018), the L^p norms utilized with these solutions are commonly used as an approximation of which filter will result in the lowest activations, and some works do in fact rely on activation-based criteria directly (Polyak and Wolf, 2015; Hu et al., 2016; Luo and Wu, 2017). Although considerable reductions in model size and complexity have been achieved with the existing pruning methods, identifying the most suitable filters to prune remains challenging (especially globally, across model layers) and requires more effective filter-importance criteria and pruning techniques.

In this paper, we address this gap and propose both, a new pruning criterion and a new iterative pruning approach that jointly lead to start-of-the-art model compression results. Here, we deviate from the established pruning concepts based on kernel norms and conjecture that the most critical filters in a layer are not necessarily the ones leading to the strongest activations, but rather the ones with very distinct activations relative to all other filters in the given layer. The main premise behind this conjecture is that such filters carry the highest amount of new information and should not be discarded due to potentially low corresponding activations. Based on this insight, we propose a new *Activation-Deviation Criterion (ADC)* in this paper that quantifies filter importance by estimating the amount of new information the filter contributes to the activation map of a given model layer. We show that ADC can be combined with standard L^p norm criteria and that such a combination convincingly outperforms the basic L^1 and L^2 approaches from the literature. Furthermore, we demonstrate that the novel criterion is easily adapted into existing solutions that address global filter importance (i.e. the importance of a specific filter in the entire network), rather than the more commonly used local importance (i.e. the importance of a specific filter in a layer), such as the recent state-of-the-art pruning approach LeGR (Chin et al., 2020). Finally, we develop a novel **I**terative **P**runing approach based on the proposed **A**ctivation **D**eviation (IPAD) criterion and evaluate it in comprehensive experiments with two sclera segmentation models and across four ocular datasets with diverse characteristics. Our experimental results show that IPAD yields highly competitive performance when compared to competing solutions from the literature and that the proposed ADC criterion leads to well pruned models capable of retaining (or even improving) the performance of the initial (over-parameterized) segmentation models.

In summary, this paper makes the following main contributions:

- We propose IPAD (**I**terative **P**runing with **A**ctivation **D**eviation), a state-of-the-art model pruning approach that iterates between: (1) pruning low-importance filters using a novel criterion for filter importance, and (2) model retraining. The main motivation for such an approach is to progressively incrementally reduce the (time/space) complexity of the initial model, while ensuring maximum performance after each pruning stage through the iterative retraining. To ensure reproducibility of our results and to facilitate further research into (mobile) ocular biometrics, we make all experimental code publicly available from sclera.fri.uni-lj.si.
- We introduce a novel **A**ctivation **D**eviation **C**riterion (ADC) for filter importance, designed for convolutional neural network pruning. ADC estimates the amount of new information a filter contributes to the activation map of the given layer. Rigorous experiments with four different datasets and two CNN models with distinct characteristics show that the proposed criterion improves on the performance of the literature-standard L^1 (Li et al., 2017) and L^2 (He et al., 2018; Chin et al., 2020) norms of the filter weights, and that it can easily be incorporated into other existing pruning solutions, such as the global-ranking-based filter pruning approach from (Chin et al., 2020).
- Supported by a comprehensive experimental analysis, we provide several new insights into the use of pruning and the proposed IPAD approach in sclera segmentation. Namely, we observe that the pruned models are in general able to retain (and in several cases even improve on) the performance of the initial unpruned models for different target FLOP (floating point operation) counts. A notable exception here are lightweight models used in cross-dataset experiments, where after the initial pruning, performance improvements are first observed. With higher FLOP reductions, on the other hand, the performance starts dropping quickly. This implies that the pruning

procedure helps identify less relevant filters and optimize the already small model architecture. However, after the weakest filters are eliminated, the removal of additional ones starts hurting the model's performance. Additionally, this observation also implies that when the models are already small, care needs to be taken when pruning them further, as this might reduce the models' generalization capabilities. Finally, we also observe that ADC most convincingly outperforms the classic L^1 and L^2 pruning as well as the other baselines when there is sufficient training data available. This is likely due to the abundance of data leading to more robust activation-based importance computation.

The remainder of this paper is structured as follows. In Section 2, we review and summarize the relevant literature, outlining the drawbacks and limitations of the existing approaches to model compression and sclera segmentation. Section 3 introduces our novel pruning criterion and elaborates on the key new ideas. Section 4 first discusses the datasets, performance metrics, and baseline models utilized for the experiments and then presents the results of our experimental evaluation. Finally, in Section 5, we summarize the main findings of the paper and conclude with some parting thoughts and directions for future research.

2. Related work

In this section, we review closely related literature most relevant to the proposed pruning approach. Specifically, we first discuss general model compression mechanisms, then elaborate on the main concepts and ideas behind model pruning, and finally review existing work in the targeted application domain, i.e., sclera segmentation. The goal of this section is to provide context for our work and further motivate the main contributions made. For a more in-depth overview of the existing literature, we refer the reader to Cheng et al. (2018), Liang et al. (2021), Choudhary et al. (2020) for comprehensive review papers on model compression and to Nigam et al. (2015); Das et al. (2013); Vitek et al. (2023) for surveys and comparative studies related to sclera segmentation.

2.1. Model compression

With the increasing availability of computational power, various areas of machine learning have been shifting towards ever larger models. While such model scaling has led to major breakthroughs, it has also introduced significant challenges when deploying the models on devices with limited computing resources, such as mobile phones or edge devices. Consequently, the field of model compression techniques has advanced in parallel with the development of deeper and more heavily parameterized models. The primary objective of model compression techniques is to reduce the memory footprint and/or computational complexity of deep learning models, making it feasible to deploy them on less capable computing hardware. Numerous solutions have emerged in this area over the years that can conveniently be grouped into three main categories, as also depicted in the taxonomy in Fig. 1, i.e.:

- *Architecture compression* techniques that aim at reducing the size (or space complexity) of the trained deep learning models while maximizing performance using pruning, knowledge distillation or dedicated model-design schemes. Techniques from this category are the most widely applicable and largely independent of the targeted deployment device and implementation frameworks.
- *Data-redundancy compression* techniques that mostly focus on decreasing the computational complexity of the deep learning models through quantization procedures, low-rank approximation schemes and weight-sharing strategies, but, depending on the mechanism used, also often reduce the space complexity. These techniques are typically dependent on the targeted processor architecture and model implementation.
- *Computation approximation* techniques that speed-up computation in deep learning models through approximation schemes based on approximate multipliers or Winograd networks.

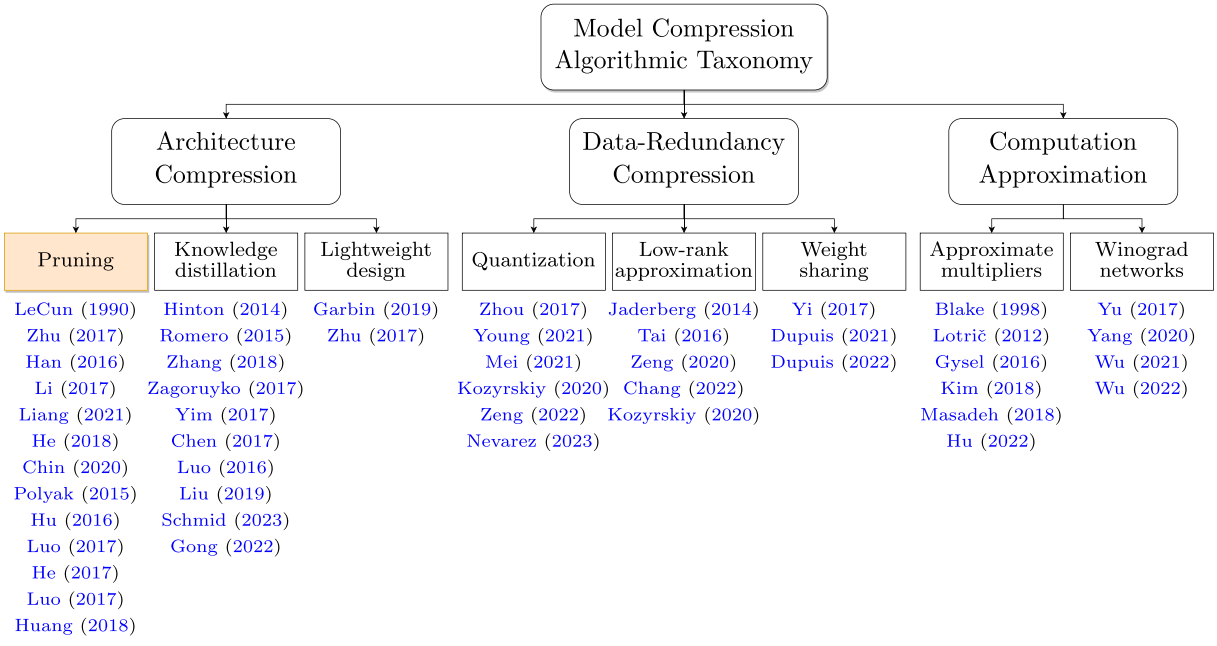

Fig. 1. Taxonomy of model compression techniques. Depending on the goal of the compression techniques, existing solutions in this area can be partitioned into three categories that target the architecture, the data representations or the mathematical operations of the deep learning models. The procedure proposed in this work falls into the group of architecture-compression algorithms based on pruning (marked orange).

Techniques from this category commonly reduce the models' computational complexity, while leaving their spatial requirements the same.

Below, we briefly review some of the recent examples from each of the three categories.

Architecture compression. One of the most notable solutions towards compressing pretrained deep learning models is *knowledge distillation* (Hinton et al., 2014). Here, the key idea is to first train a large (teacher) network and then use the outputs of the teacher network as the target (reference) outputs of a smaller (student) network. In this way, the student learns to approximate the same function as the teacher, but does so with far fewer parameters. Romero et al. (2015) generalized the distillation procedure to students that are deeper (but thinner) than the original teacher model, possibly achieving even better results than the original teacher model. Zhang et al. (2018) extended the idea of distillation by training several student models cooperatively, rather than having a single strict teacher-student relation. In Zagoruyko and Komodakis (2017), the authors incorporated attention in the distillation process, a very common concept in recent deep neural networks. Several further improvements were done in Yim et al. (2017), which introduced a novel distillation technique, useful for fast optimization, that made distillation applicable to transfer learning scenarios. Some examples of successfully applied distillation techniques to various vision problems are presented in Liu et al., 2019; Chen et al., 2017; Gong et al., 2022; Luo et al., 2016. Despite its promise, knowledge distillation has also been shown to leave significant gaps in the predictive power of the student models, even when the student should be able to match the teacher (Stanton et al., 2021).

Another popular approach from the architecture-compression category is the design and training of lightweight models from scratch. The main objective here is to construct small and lightweight models that mimic the behavior of larger and, in general, more capable models, but are trained with standard learning techniques using the available training data only. This approach has been shown to perform well in many cases (Zhu and Gupta, 2017), but in a sense predetermines a specific model size and the corresponding performance during the design stage. Additionally, it is limited in its flexibility and does not allow to effectively investigate different trade-offs between model size and performance.

The approach proposed in this papers falls into the group of pruning methods (discussed in Section 2.2) and and addresses many of the shortcomings discussed above. As we show in the experimental section, it maximizes performance, while decreasing the space complexity of the initial models and allowing for control over the complexity-vs.-performance trade-off.

Data-redundancy compression. A common approach towards reducing the data-representation redundancy in deep learning models is to rely on quantization (Zhou et al., 2017) and/or low-rank approximations (Jaderberg et al., 2014; Tai et al., 2016), both of which aim at decreasing the memory footprint of the filter-weight matrices through approximation. Quantization achieves this by replacing the floating-point representation of the weights and biases (Young et al., 2021; Nevarez et al., 2023), and possibly activations as well (Mei et al., 2021), with quantized low-bit integers (Mei et al., 2021) or even single-bit boolean values (i.e. binarization) (Zhao et al., 2017; Zeng et al., 2022). Low-rank approximation, on the other hand, focuses on optimizing each matrix or tensor as a whole, using known mathematical methods for low-rank matrix approximation, such as SVD (Chang et al., 2022) or the higher-order Tucker decomposition (Zeng et al., 2020; Kozyrskiy and Phan, 2020). Since the two methods are related, some recent work even combines them into a single cohesive method for network compression (Kozyrskiy and Phan, 2020). As noted in Jaderberg et al. (2014), low-rank approximation methods exploit the large amount of redundancy in the network's filter base. A different method of exploiting this same redundancy is weight-sharing, where certain weights are shared between various filters, thereby eliminating their redundancy (Yi et al., 2017; Dupuis et al., 2021; Dupuis et al., 2022) for a more efficient computational model. Techniques from this category are in general complementary to the architecture-compression techniques discussed above and can be used to further reduce the memory footprint of the distilled or pruned models.

Computation approximation. Techniques from this category aim to replace the computations in the deep learning models with simpler alternatives. Solutions based on approximate multiplication (Blake et al., 1998; Lotrič and Bulić, 2012; Gysel et al., 2016), for example, reduce the computational complexity of the models at a low level, replacing every multiplication operation in the network with an approximate multiplier. These approximate multipliers can result in much lower power consumption and faster execution at the cost of accuracy. However, as seen in e.g. Kim et al. (2018, 2022), the drop in accuracy is oftentimes negligible. A comprehensive comparison of approximate multipliers is available in Masadeh et al. (2018). Winograd networks, on the other hand, focus on speeding up the convolution operation by replacing it with the Winograd transformation (Yu et al., 2017; Yang et al., 2020; Wu et al., 2021; Wu et al., 2022). Similarly to data-redundancy techniques, solutions from this category represent a complementary addition, rather than an alternative for architecture compression techniques and are applicable to further simplify the targeted models.

2.2. Pruning

Model pruning is a way to reduce the memory footprint of deep neural networks by removing low-impact neurons (or filters in the case of convolutional neural networks (CNNs)) with the goal of reducing the size of the network with as little reduction in accuracy as possible. The concept of model pruning has existed since the inception of deep learning (LeCun et al., 1990) and its effectiveness has been studied extensively in Zhu and Gupta (2017). Pruning represents one of the most popular mechanism for deep neural network compression mainly due to its flexibility and the fact that it can easily be combined with other model compression mechanism, such as knowledge distillation, quantization, and approximate multiplication (Han et al., 2016).

Many different pruning approaches were introduced in the literature over the years (Liang et al., 2021), but the vast majority of modern techniques focus on quantifying the importance of each filter (or neuron) within a given network layer and then pruning away a certain fraction of the least important filters from within the analyzed layer. Here, different scoring criteria are typically utilized, but most pruning approaches rely on L^p norms (Li et al., 2017; He et al., 2018; Chin et al., 2020) of the filter weights (or some extension of this concept) as a proxy for filter importance. A notable line of research also analyzes the activations generated by the filters directly to quantify their importance (Polyak and Wolf, 2015; Hu et al., 2016; Luo and Wu, 2017) or uses other derived criteria for this task (He et al., 2017; Luo et al., 2017). Data-driven methods, such as the one proposed by Huang et al. Huang et al. (2018), have also been proposed and were shown to offer a straightforward way of controlling the model size/accuracy trade-off.

More recently, pruning approaches have also been presented that allow the estimation of filter importance across network layers and not only within the layers as discussed above. LeGR (Chin et al., 2020), for example, learns a global ranking of the filters in

the network through a data-driven approach. As a result, it is able to derive different variants of the pruned model with different complexities and accuracies. Even though the global ranking introduces another level of flexibility into the pruning process, at their core, such procedures still rely on base L^p norms to quantify the impact of the individual filters.

In this paper, we build on the research outlined above, and present a novel criterion for quantifying filter importance that can be used with standard pruning techniques for within-layer filter ranking, but also global techniques for ranking filters across different model layers. Different from existing techniques, the criterion is based on filter activations, judging filter impact based on the amount of information a given activation contributes to the overall output of the given layer. To the best of our knowledge, we are the first to consider such a differential criterion, which, as we show in the experimental section, leads to highly competitive pruning performance and is complementary to the standard L^p norm based criteria.

2.3. Sclera segmentation

Our work studies the development of lightweight models specifically for the task of sclera segmentation. An important source of information on sclera segmentation approaches and their performance is the annual Sclera Segmentation Benchmarking Competition (SSBC) (Das et al., 2015; Das et al., 2016; Das et al., 2017; Das et al., 2018; Das et al., 2019; Vitek et al., 2020b), which has been organized as part of major biometrics conferences for several years now.

Before the era of deep learning, most solutions for semantic segmentation tasks used handcrafted features and filter-based methods. Few of these were sclera-specific. Two recent examples of such methods evaluated in the scope of the SSBC competition series are the Unsupervised Sclera Segmentation (USS) approach (Riccio et al., 2017), which ranked 2nd in SSBC 2017 (Das et al., 2017), and the Sclera Segmentation using Image Properties (SSIP) techniques, which was the only entry from SSBC 2020 (Vitek et al., 2020b) not based on deep learning. Dimauro et al. (2023) presents a handcrafted approach developed for medical analysis of the sclera. Most recent sclera-segmentation solutions, including the top performers of the latest editions of SSBC, use (deep) general-purpose semantic segmentation models, usually based on the convolutional encoder-decoder (CED) architectures, such as U-Net (Ronneberger et al., 2015; Rot et al., 2020; Vitek et al., 2020b; Lv et al., 2022; Wang et al., 2022; Das et al., 2022), SegNet (Badrinarayanan et al., 2017; Das et al., 2017; Rot et al., 2020; Vitek et al., 2020a; Rot et al., 2018), ScleraSegNet (Wang et al., 2019; Das et al., 2019), RefineNet (Lin et al., 2017, 2018, 2020,a), and DeepLab (Chen et al., 2018; Vitek et al., 2020b).

Such models perform quite well, but have a large number of parameters and complex architectures, which makes them expensive in terms of both computation and memory requirements. We aim to implement a lightweight model that comes as close as possible to the performance of these larger, heavily parameterized models, but exhibits only a fraction of the memory footprint and a significantly reduced FLOP count.

3. Methodology

The main contribution of this work is the IPAD (**I**terative **P**runing with **A**ctivation **D**eviation) pruning approach, which iteratively reduces the computational complexity of deep learning models by pruning away the lowest-impact filters, identified through a novel activation-deviation criterion (ADC), as shown in Fig. 2. We study IPAD in the context of sclera-segmentation models. In this section, we first present a high-level overview of the proposed approach and discuss its characteristics, then introduce the novel pruning

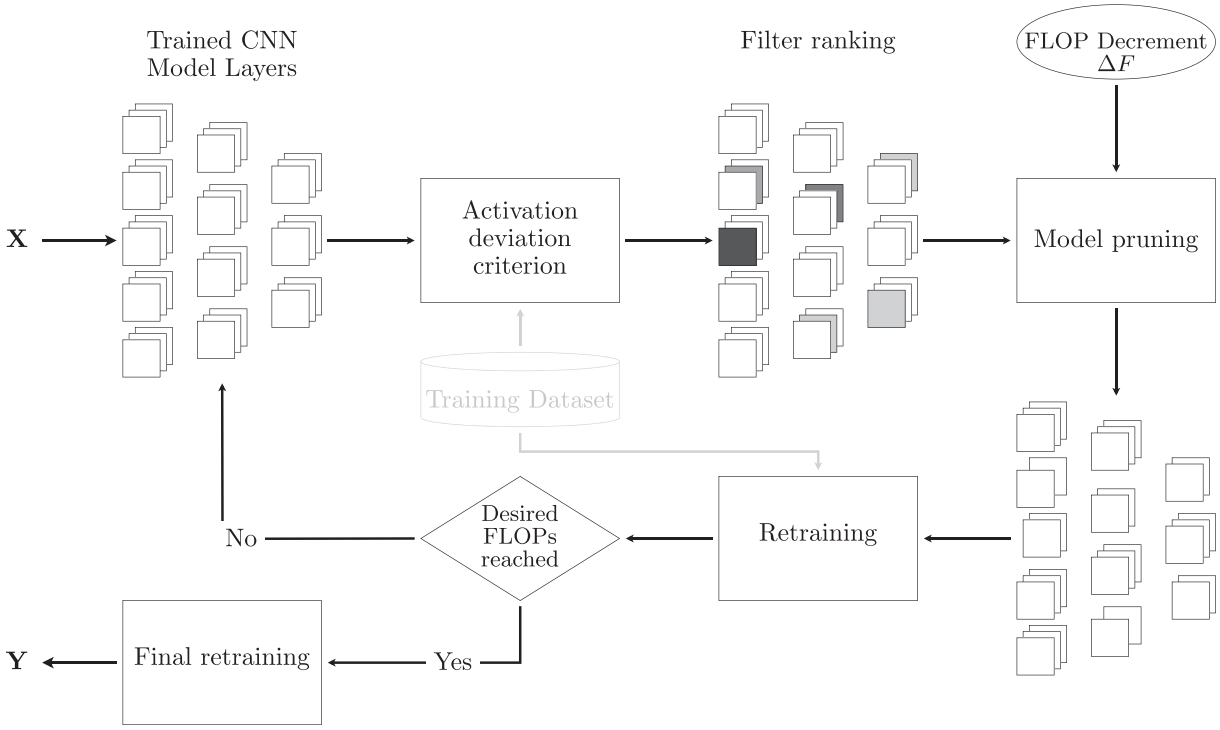

Fig. 2. Overview of the Iterative Pruning with Activation Deviation (IPAD) approach. IPAD takes an existing deep learning model as input (top left) and then iteratively prunes the lowest-impact filters to produce a simpler model with reduced complexity (bottom right). During each iteration, IPAD reduces the number of FLOPs of the model by a predefined decrement ΔF. After each pruning step, the model is retrained to ensure optimal performance given the current topology. The procedure is repeated until the desired FLOP count is reached. At the core of IPAD is a novel Activation-Deviation Criterion (ADC) that together with the standard L^p norm criterion drives the pruning procedure.

criterion and finally recapitulate on the entire approach through a step-by-step summary.

3.1. Iterative Pruning with Activation Deviation (IPAD)

Assume a well-trained (overparameterized) deep-learning model $\mathbf{X}$ with N layers and l_i filter kernels $\{\mathbf{K}_n\}_{n=1}^{l_i}$ in each layer, where $i \in \{1, \cdots, N\}$. The goal of the IPAD procedure is to produce a (optimal) pruned model $\mathbf{Y}^*$ based on the following constrained maximization problem (He et al., 2017):

$$\mathbf{Y}^* = \underset{\mathbf{Y} \in \mathcal{Y}}{\operatorname{argmax}} \{P(\mathbf{Y})\}, \quad \text{s.t.} \quad \text{FLOPs}(\mathbf{Y}) \leqslant p \cdot \text{FLOPs}(\mathbf{X}), \tag{1}$$

where $\mathcal{Y}$ is the set of deep models with a subset $\mathcal{K}$ of $\mathbf{X}$'s filters (i.e., $\mathcal{K} \subset \{\mathbf{K}_n\}_{n=1}^{l_i}$), P is the scoring function used to evaluate the performance of the model, and $p \in [0, 1]$ is the targeted fraction of the floating point operations (FLOPs) to be retained. Thus, the overall objective is to identify a pruned model $\mathbf{Y}$ that maximizes performance, while needing fewer FLOPs than the initial model $\mathbf{X}$ for processing a given input. Because we are targeting segmentation models in this work, P is defined as a function that returns the average Intersection-over-Union (IoU) over the available training data (Vitek et al., 2020a; Rot et al., 2018; Lozej et al., 2018).

As illustrated in Fig. 2, IPAD approaches the optimization process in Eq. (1) through an iterative procedure that prunes the filters of $\mathbf{X}$ gradually in increments that correspond to a predefined reduction ΔF in the FLOP count of the model. In each step of the optimization process, the least important filters are pruned and the model is retrained for a fixed number of epochs. This process is repeated until the desired model complexity is reached, at which point the final (pruned) model is retrained once more until convergence. Because the filter importance is determined for each layer separately, we impose an upper limit on how many filters can get pruned from a given layer to maintain the desired overall model architecture, similarly to competing approaches from the literature (Shang et al., 2022).

The main motivation for the iterative pruning process used with IPAD is twofold:

- **Complexity-performance trade-off:** The iterative nature of IPAD and incremental removal of filters corresponding to a computing budget of ΔF FLOPs, allows for fine-grained control of the complexity-performance trade-off ensured by the pruned models. Because the model is retrained after each pruning step, the loss in model performance due to the pruning-induced model reparameterization is explicitly minimized. This characteristic represents a unique aspect of IPAD not available with the majority of competing techniques.
- **Global relevance:** IPAD quantifies filter importance in a local manner, i.e., separately for each model layer, similarly to the majority of existing pruning techniques (Liang et al., 2021). Since filters are pruned locally, the fixed-epoch retraining step (conducted at each iteration) updates the remaining filters and, in a sense, readjusts their global importance. Thus, the iterative procedure contributes towards the global relevance of the local (iterative) pruning process and addresses the limitations of existing pruning approaches that are typically of a local nature.

The key component for successful and well-performing model pruning is the criterion utilized for determining filter importance. For IPAD, we develop a novel criterion for this task that relies on filter activations rather than kernel norms and is presented in detail in the next section.

3.2. Novel pruning criterion

One of the most important parts of any pruning procedure is the selection of the low-impact neuron(s) or (in our case) filter(s) that can be removed with minimal impact on performance. A common way of quantifying the filter importance is to compute the L^1 or L^2 norm of the filter weights (Li et al., 2017; He et al., 2018; Chin et al., 2020), i.e.:

$$\psi_{w1}(\mathbf{K}_n) = \|\mathbf{K}_n\|_1 = \sum_{i=1}^{h_k} \sum_{j=1}^{w_k} |\mathbf{K}_n(i,j)| \qquad \text{or} \tag{2a}$$

$$\psi_{w2}(\mathbf{K}_n) = \|\mathbf{K}_n\|_2 = \sqrt{\sum_{i=1}^{h_k} \sum_{j=1}^{w_k} (\mathbf{K}_n(i,j))^2}, \tag{2b}$$

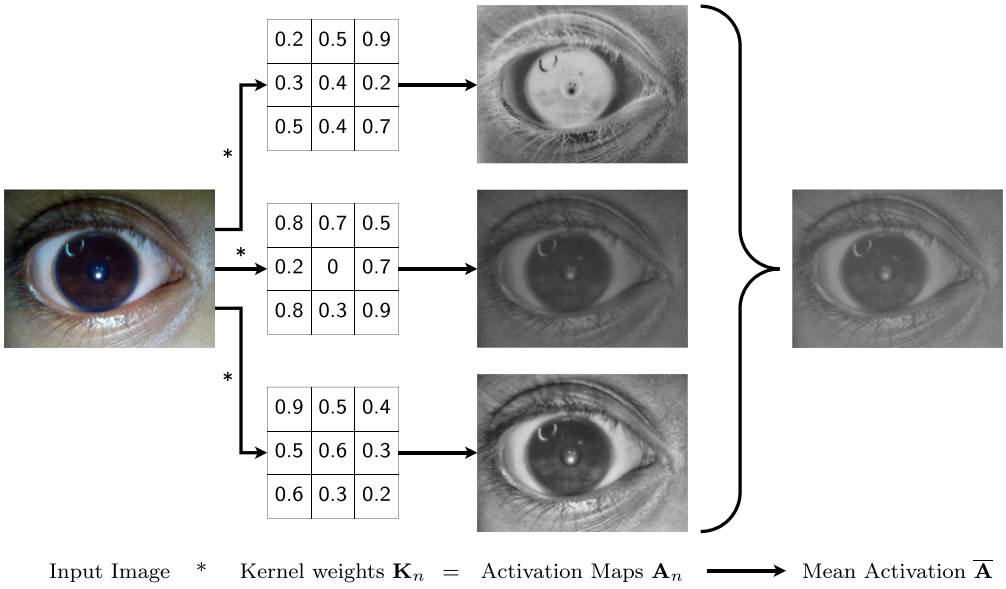

Fig. 3. Computation of the activation maps $\mathbf{A}_n$ from an input image within a given layer. The activation maps are used to compute the mean activation map of the whole layer $\overline{\mathbf{A}}$. The figure is illustrative.

where $\psi_{wx} : \mathbb{R}^{h_k \times w_k \times c} \rightarrow \mathbb{R}$ is the standard weights-based criterion, $\mathbf{K}_n \in \mathbb{R}^{h_k \times w_k \times c}$ is the weight matrix of n-th filter in the given layer, h_k and w_k are the kernel height and width, and c is the number of channels, i.e., the kernel depth. The main assumption here is that filters with small weights, and consequently, small norms, contribute little to the outputs/activations of the model and, in turn, can be removed from the model.

However, because such weight-based criteria serve only as proxies for the expected filter activations (He et al., 2018), which correlate more closely with filter importance, we propose a novel Activation-Deviation Criterion (ADC) in this section that predicts filter impact directly from the activations produced over some training dataset. To derive our activation-based criterion, we begin with the activation maps $\mathbf{A}_n \in \mathbb{R}^{h \times w \times l_i}$ of the model, presented in Fig. 3, where h and w are the height and width of the activation maps, respectively, and l_i is the number of output channels, i.e. the number of filters in the layer. Note that in the actual implementation, the criterion is computed over batches of input images to increase robustness, however for a simplified illustration we rely on a single input image throughout the entire derivation. To capture the filter importance, ADC quantifies the deviation of each filter's activation map from the overall mean activation of all filters $\overline{\mathbf{A}}$ within a given layer. Thus, ADC first computes the mean activation $\overline{\mathbf{A}}$, as illustrated in Fig. 3:

$$\overline{\mathbf{A}} = \frac{1}{l_i}\sum_{n=1}^{l_i}\mathbf{A}_n, \tag{3}$$

where the n-th activation map $\mathbf{A}_n$ is computed by convolving the input with the n-th kernel $\mathbf{K}_n$ of the given layer. The corresponding deviation from the mean activation is then determined as:

$$\mathbf{D}_n = \mathbf{A}_n - \overline{\mathbf{A}}. \tag{4}$$

These deviations $\mathbf{D}_n$ serve as a measure of how much new information the n-th filter brings to the overall activation map of a given model layer. The computation procedure is illustrated in Fig. 4.

Finally, to obtain a scalar measure of the importance of a filter, ADC uses the standard L^1 or L^2 norms by applying them to the difference matrices computed in the previous step:

$$\psi_{a1}(\mathbf{A}_n) = \frac{\|\mathbf{D}_n\|_1}{h \cdot w} = \frac{\sum_{i=1}^{h}\sum_{j=1}^{w}|\mathbf{D}_n(i,j)|}{h \cdot w} \quad \text{or} \tag{5a}$$

$$\psi_{a2}(\mathbf{A}_n) = \frac{\|\mathbf{D}_n\|_2}{h \cdot w} = \frac{\sqrt{\sum_{i=1}^{h}\sum_{j=1}^{w}(\mathbf{D}_n(i,j))^2}}{h \cdot w}, \tag{5b}$$

where ψ_{ax} is the ADC criterion and x defines the norm type. Note that the standard weight-based pruning criteria from Eqs. 2a, 2b operate on $h_k \times w_k$ kernels, whereas the ADC criterion operates on $h \times w$ activations maps, where typically $w \gg w_k$ and $h \gg h_k$. For IPAD, we therefore define a combined criterion $\psi(n)$ to have a comprehensive and complementary description of the n-th filter importance, i.e.:

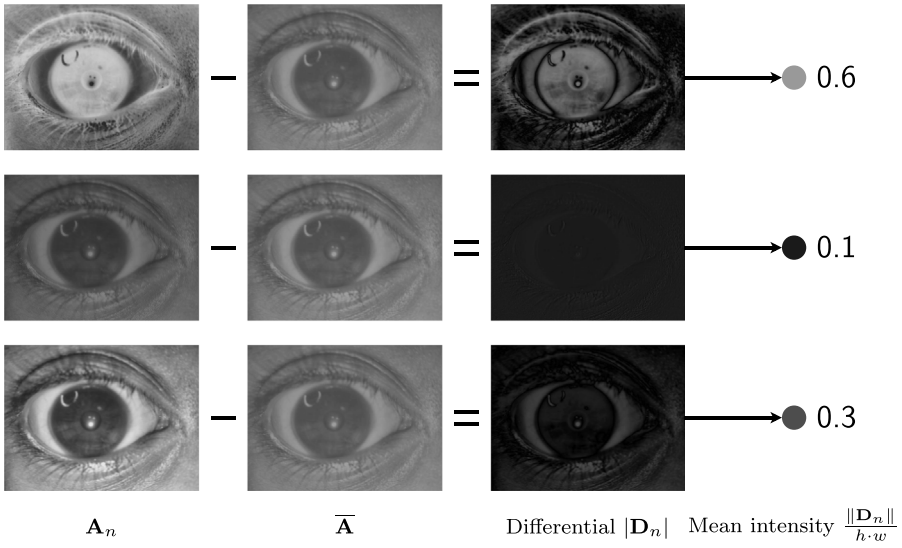

Fig. 4. Computation of the difference matrices $\mathbf{D}_n$. The difference matrices capture the amount of new information each filter's activation brings to the overall layer's activation map. The scalar filter importance is then computed as the mean intensity of each difference matrix. Note that for illustration purposes, we show the absolute values of $\mathbf{D}_n$.

$$\psi(n) = \alpha \cdot \psi_w(\mathbf{K}_n) + (1-\alpha) \cdot \psi_a(\mathbf{A}_n), \tag{6}$$

Table 1
High-level characteristics of the four datasets used in the experiments. The datasets differ in terms of image resolution, acquisition devices used, gaze directions and blur, but also in the amount of data available.

Dataset	#Images	#IDs	#Eyes	Resolution [px]	Sources of Variability[†]
SMD (Das, 2017)	500	25	50	3264 × 2448	BL, CN
SLD (Vitek et al., 2023)	108	27	54	3264 × 2448	BL, CN
SVBPI (Vitek et al., 2020a; Rot et al., 2020)	1858	55	110	3000 × 1700	GZ, BL
MOBIUS (Vitek et al., 2020b; Vitek et al., 2023)	3542	35	70	3000 × 1700	MD, CN, GZ, BL

[†] GZ - gaze, BL - blur, CN - acquisition condition, MD - mobile device.

where α is the weighting parameter that determines the trade-off between the two criteria. In each IPAD iteration, we prune the filter with the lowest importance.

3.3. IPAD pseudocode

Using the newly proposed criterion, the complete IPAD pruning approach is implemented in accordance with the pseudocode provided in Fig. 1.

Algorithm 1.

Algorithm 1: IPAD pruning method.

Input: $\mathbf{X}$, training_data

Hyperparameters: p, FLOPs_decrement

Output: pruned model $\mathbf{Y}^*$

```
/* Initial model training                                          */
train(X, training_data)

/* IPAD method                                                     */
target_FLOPs ← p · FLOPs(X)
while FLOPs(X) > target_FLOPs do
    for filter ∈ X.filters do
        filter.compute_importance()                    // Eq. (6)
    end
    Sort X.filters in ascending order by importance
    removed ← 0
    while removed < FLOPs_decrement do
        removed ← removed + FLOPs(X.filters[0])
        prune(X.filters[0])
    end
    train_fixed_epochs(X, training_data)
end
train(X, training_data)
Y* ← X
return Y*
```

4. Experiments and results

In this section, we present experiments conducted to evaluate IPAD and the proposed filter-importance criterion. We start the section with a description of the experimental datasets, performance metrics, and hyperparameters used for the evaluations and then proceed with the presentation and discussion of the results.

4.1. Datasets

Four datasets were used for our experimental work, two of which (MOBIUS and SBVPI) were collected at the University of Ljubljana and are publicly available for research purposes from sclera.fri.uni-lj.si. The remaining two (SMD and SLD) are external datasets but are also publicly available on request (Vitek et al., 2023). Details on the four datasets are given below and their key characteristics are summarized in Table 1.

- **SBVPI** (Rot et al., 2020; Vitek et al., 2020a) is a dataset of 1858 images of 55 subjects (i.e. 110 eyes) with corresponding hand-crafted markups of the sclera and the periocular region. A subset of roughly 130 images is also annotated with the sclera vessels, pupil, iris, canthus, and eyelashes. The samples in the dataset were captured in laboratory conditions with a DSLR camera, and they are therefore high-quality high-resolution ocular images acquired in well-lit conditions. The images come

with labels for the corresponding subject ID, eye (left/right), and gaze direction (left/right/up/straight). The dataset also contains additional subject information, namely age, gender, and eye colour. We show some sample images, as well as region markups from the dataset in Fig. 5(a).

- The **MOBIUS** dataset (Vitek et al., 2020b; Vitek et al., 2023), shown in Fig. 5(b), is a mobile ocular dataset of almost 17000 ocular images belonging to 100 subjects (i.e. 200 eyes). Its segmentation subset – which contains 3542 images from 35 subjects (70 eyes) – comes bundled with manually crafted ground truth markups for the sclera, pupil, iris, and periocular region. The images in the dataset were acquired in different capturing conditions: using 3 different mobile phones (Sony Xperia Z5 Compact, Apple iPhone 6s, Xiaomi Pocophone F1), in 3 different lighting conditions (natural lighting, indoor lighting, unlit indoor room), and with 4 different gaze directions (left/right/up/straight). The dataset additionally contains some deliberately unusable ("bad") images, which contain image noise (such as motion blur, obstructions, etc.) and are intended as negative samples in quality control. Since our experiments do not include the study of quality assessment, we discard the bad images to obtain the final dataset of 3475 images. All capturing conditions, as well as the corresponding subject ID and eye (left/right), are labelled in the image names. Additionally, the dataset contains rich subject metadata, including information about their age, gender, eye colour, dioptres and other medical conditions, allergies, whether they smoke, and whether they wore lenses or used eyedrops at the time of the image acquisition.
- The **SMD** (Das, 2017) and **SLD** (Vitek et al., 2023), shown in Figs. 5(c), 5(d), are external datasets, obtained and used with the permission of their author. SMD is a dataset of 500 images from 25 individuals (i.e. 50 eyes), captured using a mobile camera in different lighting conditions. It has been used in several SSBC competitions (Das et al., 2019; Vitek et al., 2020b). SLD is a smaller dataset of 108 images from 27 individuals (54 eyes) captured by a mobile camera under different gaze directions, which was developed primarily for sclera liveness detection. It was also utilized for the recent exploration of demographic and algorithmic bias in sclera segmentation methods (Vitek et al., 2023).

We split all of our datasets into a training set (used for training the segmentation models), validation set (used for early stopping and hyperparameter selection), and testing set (used for evaluation) in a 70/20/10% split. The cross-dataset experiments (see Section 4.5) instead use a 70/30% split on SMD for training and validation data, and use the entire SLD dataset for evaluation (i.e., performance reporting). Depending on the ground-truth annotations available with the four datasets we conduct either 2-class (sclera vs. the rest) or 4-class (sclera, iris, pupil, and background) segmentation experiments.

4.2. Performance metrics

The primary performance indicator used throughout our experiments as a measure of model accuracy is *IoU* (Intersection-over-Union), a standard measure in the field of semantic segmentation (Vitek et al., 2023; Vitek et al., 2020a). For the 2-class segmentation task we use the *IoU* of the positive class (i.e. sclera), defined as:

$$IoU = \frac{|P \cap T|}{|P \cup T|}, \tag{7}$$

where P is the set of the pixels predicted to belong to the sclera by the model and T are the actual sclera pixels, whereas for the 4-class ocular segmentation problem we use *mIoU* (mean intersection-over-union), which is defined as:

$$mIoU = \frac{1}{C}\sum_{c=1}^{C} IoU_c, \tag{8}$$

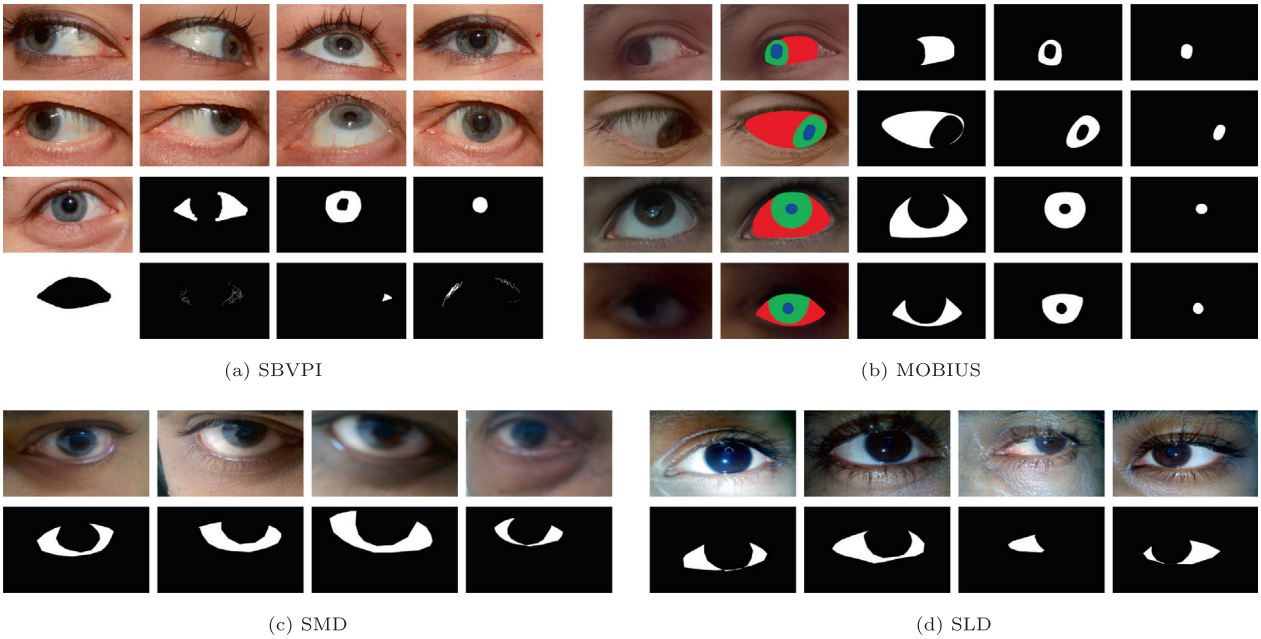

Fig. 5. Sample images from the four datasets used in the experiments. The top two rows of (a) show different gaze directions and eye colours present in SBVPI, while the bottom two rows show (left-to-right and top-to-bottom) a sample and the ground truth markups of the: sclera, iris, pupil, periocular region, scleral vessels, medial canthus, and eyelashes. (b) shows four samples from MOBIUS captured in different capturing conditions. The image in the first row was captured in natural sunlight, the second row in a well-lit indoor room, the third in an unlit room, while the last row displays an intentionally unusable ("bad") image intended for quality control. Each row additionally contains the corresponding ground-truth multi-class markups and the individual masks for the sclera (red), iris (green), and pupil (blue). Finally, (c) and (d) show varied images from SMD and SLD, respectively, along with their corresponding sclera ground truth markups.

where $c \in \{1, \ldots, C\}$ is the class index, P_c is the set of pixels recognized by the model as class c, T_c is the set of pixels actually belonging to class c, and $IoU_c = \frac{|P_c \cap T_c|}{|P_c \cup T_c|}$ is the class-specific intersection-over-union.

4.3. Baseline models and training procedure

We use two distinct segmentation models for the experiments to explore different aspects of the proposed IPAD pruning procedure and ADC criterion, i.e., RITnet (Chaudhary et al., 2019) and U-Net (Ronneberger et al., 2015):

- **RITnet:** This is a fairly lightweight model and was shown to be very effective for the task of ocular segmentation in the 2019 OpenEDS challenge (Garbin et al., 2019). Having been designed specifically for multi-class (sclera, pupil, iris, background) segmentation problems, it is also well suited for our intended purposes. From an architectural point of view, RITnet is a lightweight (16 GFLOPs, 250000 parameters) convolutional encoder-decoder (CED) model inspired by the DenseNet (Huang et al., 2017) architecture. The encoder of the model consists of 5 Downsampling-Blocks, which contain 5 convolutional layers each, followed by an average pooling layer. The decoder consists of 4 Upsampling-Blocks that upsample the output of the encoder back to the original image resolution via the nearest-neighbor method. The Upsampling-Blocks each contain 4 convolution layers and skip-connections to their respective Downsampling-Block. The model serves in our experiments to demonstrate the performance of the proposed pruning procedure on an already compact segmentation model.
- **U-Net:** This model represents a go-to solution for many image-to-image translation tasks, including semantic image segmentation (Rot et al., 2018; Vitek et al., 2023). Similarly to RITnet, U-Net also features an encoder-decoder architecture, where the encoder begins with a Double-Convolution-Block (DCB), which consists of two pairs of convolutional layers and batch normalization with ReLU activation. The encoder then continues with 4 Downsampling-Blocks, each of which consists of a max-pooling layer followed by a DCB. The decoder contains 4 Upsampling-Blocks, each of which consists of a bilinear upsampling layer followed by a DCB, and a final 1×1 convolution that ensures the number of the model's output channels matches the desired number of classes. The Upsampling-Blocks again contain skip-connections to their respective Downsampling-Blocks. The model configuration used in our experiments has a total of 17.3 million parameters with roughly 160 GFLOPs ($10\times$ as many as RITnet) and is used to demonstrate the characteristics of IPAD with a heavily parameterized and computationally more complex model topology.

We train both models using the same training process and loss to ensure a fair comparison. Specifically, we utilize the learning objective proposed in Chaudhary et al. (2019), which was designed specifically with ocular segmentation in mind, and is defined as:

$$L_R = l_{CE}(\lambda_1 + \lambda_2 l_E) + \lambda_3 l_{GD} + \lambda_4 l_S, \tag{9}$$

where l_{CE} is the pixel-wise cross-entropy loss, which penalizes incorrect pixel classifications and is a standard loss in semantic segmentation, but is primarily designed for use with balanced classes; l_E is the Canny edge loss, which maximizes the accuracy of the detection of edges between regions by weighting the pixels by their distances to the nearest two image segments; l_{GD} is the generalized dice loss, which ensures stable gradients in the case of imbalanced classes (which are common for ocular segmentation problems) by weighting the dice score by the squared inverse of the class frequency; l_S is the surface loss, which is based on the contour distances and aims to preserve smaller areas ignored by the previous two losses. For further details about the loss components and the selection of the λ parameters we refer the reader to the original paper (Chaudhary et al., 2019).

We train the models for 200 epochs, which was determined to be sufficient for proper convergence of both models. Additionally, we use a separate set of data samples for validation (distinct from the training and testing data) to implement early stopping criteria that help avoid overfitting. Specifically, we consider the model to have converged and end the training early if the loss on the validation data does not improve in 10 consecutive epochs. The final retraining after the pruning procedure is carried out in the same manner, using the same loss function, on the smaller model with pruned filters. The brief retraining during pruning also follows the outlined procedure, but is only executed for 5 epochs with no stopping criteria. The number of epochs for the retraining during the pruning iterations was determined through preliminary experiments by selecting a trade-off between IPAD runtime complexity and impact on the final accuracy score.

For the learning process, we use the Adam optimizer with a learning rate of 0.001. For the weight parameters λ from Eq. (9), we use the optimal values advocated in the original paper (Chaudhary et al., 2019) ($\lambda_1 = 1, \lambda_2 = 20, \lambda_3 = 1 - \lambda_4$, and $\lambda_4 = max\left(1 - \frac{epoch}{125}, 0\right)$), and prune filters away in increments corresponding to $\Delta F = 1$ GFLOP. During pruning, we limit the number of filters that can be pruned from each layer or each of RITnet's blocks to 75%. When we prune away a filter, we also adjust the dimensions of the subsequent filters and batch normalization layers that depend on it and remove its corresponding bias. We never prune batch normalization layers directly. For the experiments, all images are resized to 640×400 pixels. The training and experiments are conducted on various graphic cards, specifically GeForce TITAN V, GeForce RTX A5000, and several GeForce RTX 3090s and GeForce RTX 2070s.

4.4. Benchmark methods

We compare IPAD to several pruning methods from the literature, including the standard L^1 (Li et al., 2017) and L^2 norm-based pruning methods (He et al., 2018; Chin et al., 2020), and the global LeGR approach from Chin et al. (2020). Additionally, we also implement *uniform* and *random* pruning procedures for reference comparisons, neither of which uses any data/model derived criterion to determine filter importance but instead prunes filters at random in slightly different ways. Thus, we compare IPAD to the following pruning methods:

- **LeGR** (Chin et al., 2020) is a state-of-the-art pruning method from the literature which uses a layer-local pruning criterion (such as L^1 or L^2 weight norms or our criterion), but additionally learns scale and shift parameters for each layer to facilitate global filter comparisons. In this way the method obtains a global ranking of the model's filters and different from most competing solutions allows for efficient global model pruning.
- Weights-based $\boldsymbol{L^1}$ (Li et al., 2017) and $\boldsymbol{L^2}$ (He et al., 2018; Chin et al., 2020) pruning are classical pruning methods from the literature, which use the L^1 and L^2 norms of the kernel weights as the criteria for filter importance. We apply these criteria in our experiments using the same pruning/retraining strategy as implemented for the proposed IPAD method to ensure a fair comparison.

- **Uniform** pruning (Liu et al., 2018) iterates over the layers in the network and prunes a single randomly selected filter from each layer of the network. After reaching the final layer of the network, the procedure starts over with the first layer and continues in this manner until the desired amount of FLOPs for the given iteration ($\Delta F = 1$ GFLOP in our implementation) is removed.
- **Random** pruning (Li et al., 2022) is the simplest possible pruning procedure that takes a single randomly selected filter from the entire network and prunes it. It then repeats this process until the desired amount of FLOPs is removed.

Finally, we also compare the pruned models with the original **unpruned** version, i.e., the result of the initial model training (see Fig. 1). In all graphs and figures, the results for the pruned models are reported at 50% of the FLOP count of the unpruned model (i.e. roughly 8 GFLOPs for RITnet and 80 GFLOPs for U-Net), unless specified otherwise. It is important to note that we count *one multiplication and one addition* as a single operation when reporting the computational complexity of the models, as modern processor architectures implement such a pair as a single MAC (multiply-accumulate) instruction (IEEE Standard, 2008).

4.5. Results

In this section, we present the results of our experimental assessment. We compare the pruning methods on RITnet and U-Net in 4 different settings:

- **High-quality sclera segmentation** using the laboratory-quality images from the SBVPI dataset. This setting allows us to test the models' performance in settings where we have a high degree of control over the environmental factors, such lighting. The SBVPI dataset contains the high-quality images captured in ideal conditions and is large enough to facilitate effective training of our model. Using this scenario, we therefore explore the impact of pruning in ideal conditions, i.e., with plentiful high-quality and well annotated training data, which should help both small and large models achieve relatively good performance.
- **Limited-data in-the-wild sclera segmentation** using the SMD dataset, which contains a smaller number of images, all captured by a mobile camera in real-world conditions. This scenario tests the performance of the segmentation models in more unconstrained, real-world environments, such as the ones encountered in mobile-phone unlocking tasks. With the smaller SMD dataset, the segmentation models additionally have a lower number of training images available, introducing another source of difficulty. The small amount of training data can cause larger models to overfit, and it has been shown (Brigato and Iocchi, 2020) that smaller networks can actually outperform larger ones when lacking training samples, making this setting very interesting for the investigation of the proposed pruning procedure.
- **Cross-dataset sclera segmentation**, where the segmentation models are trained on images from SMD and evaluated on SLD – a small dataset of ocular images acquired in real-world conditions, intended for sclera liveness detection. With this experiment, we evaluate the ability of the segmentation models to generalize and adapt to new data samples that are significantly different from anything the model saw during training. Studying the impact of pruning in this scenario is particularly interesting, as it contains two converse problems: (*i*) low amount of training data (with which, as described above, smaller models can actually perform better), and (*ii*) generalization to distinct unseen data, where it is known (Neyshabur et al., 2015; Novak et al., 2018) that more complex networks tend to generalize better.
- **Four-class ocular segmentation** on the images from the MOBIUS dataset, which were captured using three different mobile cameras in three different real-world lighting conditions – outdoor natural light, indoor lighting, poorly lit indoor room. This scenario avails the segmentation models of a plethora of training images but again places them into an unconstrained real-world environment, in the more challenging task of four-class segmentation. With this experiment, we explore the adaptability of our pruning criterion and method to different tasks, while also evaluating the models in a hybrid setting with (*i*) a large number of training examples (similar to the high-quality setting), but (*ii*) worse capturing conditions (similar to the in-the-wild setting).

We first demonstrate the superior performance of our criterion relative to the classic weights-based criterion in Sub Section 4.5.1, where we compare the performances of the two side-by-side. Next, we study how the pruning affects the model's performance, as we prune more and more filters in Sub Section 4.5.2, in which we compare the performance of the models at different FLOP counts using each of the pruning procedures. In Sub Section 4.5.3 we investigate the importance of the proper choice of the α parameter. The ablation study in Sub Section 4.5.4 explores the removal of the weights-based and activation-based components, as well as the α selection process, and finally also shows: (*i*) the impact of removing 1×1 pruning from our pruning procedure entirely, and (*ii*) the impact of not using our criterion on the 1×1 filters.

4.5.1. Comparison with previous work

In the first set of experiments, we look at the performance differences that arise from the use of the proposed filter-importance criterion $\psi(\cdot)$ from Eq. (6) that forms the basis for IPAD. We note that the proposed criterion extends the standard L^1 and L^2 (filter-weights) norm criteria to also consider activation deviations when determining filter importance, and, therefore, includes the standard criteria as a special case when $\alpha = 1$. Furthermore, all existing baseline techniques can also be implemented using the iterative procedure, introduced for IPAD. In the experimental evaluations, we, therefore, compare all considered baseline pruning methods side-by-side first using:

- the optimal balancing weight α (denoted "*IPAD* ($\alpha = opt.$)" in the figures) that resulted in the best performance (i.e., highest IoU scores) on the validation data, where $\alpha \neq 1$, and
- only the filter-weights norm criterion denoted as "*Weights only* ($\alpha = 1$)" in the figures.

In all visualizations, the orange bars correspond to the use of the proposed (combined) criterion $\psi(\cdot)$ with different configurations of the IPAD method (L^1 vs. L^2 vs. L^1 or L^2 LeGR), the pink bars correspond to the respective implementations with the weight-based criterion only, and the blue bars correspond to the reference methods. The results of the pruned models are reported at 50% of the original unpruned model's FLOP count. The reader is referred to the Appendix for the full table of results and a statistical analysis of the impact of our criterion.

High-quality sclera segmentation: On the high-quality data of SBVPI both segmentation models achieve high *IoU* scores, as shown in Fig. 6. In 7 out of the 8 pairs, the proposed criterion outperforms the standard weights-based criterion. Also note that the best result overall for both models (RITnet and U-Net) is achieved using the proposed combined criterion. Additionally, while the standard criterion is in some cases outperformed even by random and uniform pruning (in terms of final IoU score), this is never the case for the newly proposed criterion. Finally, note that both criteria on both

considered segmentation models consistently outperform the original unpruned model, despite having only 50% of its computational complexity, which is also in line with the fact that RITnet (a lightweight model designed specifically for ocular segmentation) outperforms U-Net (a more general larger model), despite having only 10% of its FLOP count. Overall, we observe that the proposed criterion leads to both better and more consistent results than the classic weights-based criterion alone.

Limited-data in-the-wild sclera segmentation: In this more challenging setting, the models achieve noticeably lower *IoU* scores overall, as expected due to the more challenging images and lower number of training samples available. As can be seen from Fig. 7, the proposed combined criterion outperforms the standard criterion, with 6 out of the 8 pairs and the best overall result for both segmentation models again being achieved by the combined criterion $\psi(\cdot)$. Additionally, while the standard criterion is in some cases outperformed even by random and uniform pruning, this only happens once with our criterion in the case of L^2 pruning on U-Net, where both criteria performed poorly (worse than uniform pruning). Finally, note that both criteria with both segmentation models again consistently lead to smaller models that even outperform the original unpruned model. This result suggests that both initial models (RITnet and U-Net) are over-parameterized given the studied segmentation task and that (after pruning) the retraining results in more capable segmentation networks, whose complexity better suits the targeted task.

Cross-dataset sclera segmentation: In the cross-dataset experiments, the performance of both segmentation models is overall worse (with lower IoU scores) than in the previous experiments, where training and testing were conducted on (disjoint) images coming from the same dataset, as shown in Fig. 8. With 5 out of the 8 pairs of pruning methods, the combined filter-importance criterion outperforms the standard weights-based criterion and the best overall result for both segmentation models is again achieved with the proposed criterion $\psi(\cdot)$. However, we do observe that the results are less consistent than with the within-dataset experiments due to the more challenging setting. We observe that in the cross-dataset setting, the standard criterion is outperformed by random and uniform pruning in 5 out of 8 cases, while this only happens in only 2 out of 8 cases with the proposed filter-importance criterion. Additionally, RITnet in this experiment exhibits far better performance in its unpruned state, only being outperformed through the use of the proposed criterion in 2 of the 4 cases and never by the standard criterion. Overall, our criterion maintains its superiority over the classic weights-based criterion and in general leads to better performing pruned models, but is still less consistent than in the previous experiments due to the more challenging task, in which filter pruning has a bigger impact on performance.

Four-class ocular segmentation: In this multi-class problem, the segmentation models achieve decently high *mIoU* scores, as shown in Fig. 9. With 7 out of the 8 pairs, the proposed criterion outperforms the standard criterion and once again leads to the best overall result for both models (RITnet and U-Net) in terms of segmentation performance. Additionally, while the standard criterion is in some cases outperformed even by random and uniform pruning, this is never the case with our criterion. Finally, note that both criteria on both segmentation models consistently outperform the original unpruned model despite the simpler architecture and reduced FLOP count. Similarly as with the experiments discussed above, the proposed combined filter-importance criterion once again performs better and more consistently than the classic weights-based criterion.

4.5.2. Performance across different complexities

In the previous section, we observed that the pruned models very often outperform the corresponding original (unpruned) models despite their much lower computational complexity in terms of FLOP count. This observation can be attributed to the fact that, given the task at hand, both considered segmentation models are over-parameterized. This prompts us to also look at the impact of different model complexities on the final segmentation perfor-

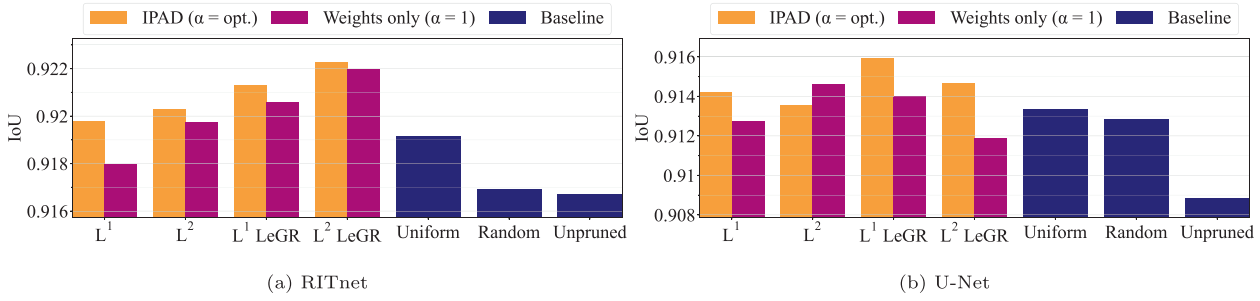

Fig. 6. Impact of the proposed filter-importance criterion $\psi(\cdot)$ on SBVPI. The segmentation performance in the *high-quality sclera segmentation* setting is reported for the pruning methods implemented with the proposed criterion vs. the standard weights-based criterion normally used in the literature.

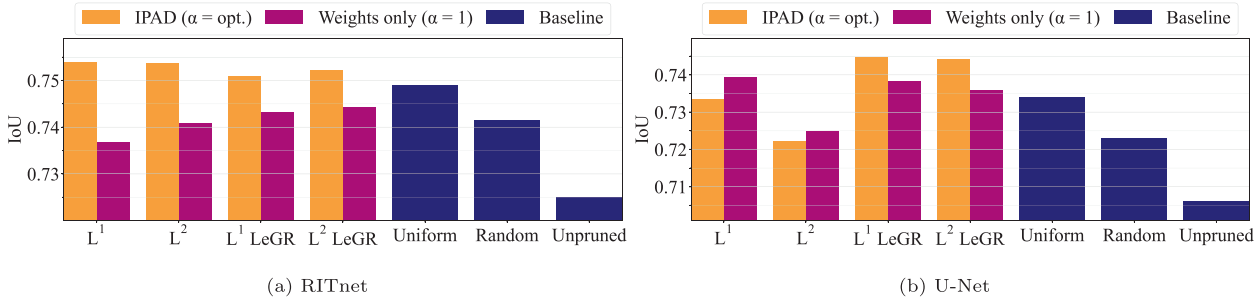

Fig. 7. Impact of the proposed filter-importance criterion $\psi(\cdot)$ on SMD. The segmentation performance in the *limited-data in-the-wild sclera segmentation* setting is reported for the pruning methods implemented with the proposed criterion vs. the standard weights-based criterion normally used in the literature.

mance achieved by the pruned RITnet and U-Net models. To study the behavior of the proposed pruning procedure with different target FLOP counts and explore the segmentation performance of the pruned models in this experimental series, we set 3 different targets at 25%, 50% and 75% of the initial FLOP count of the unpruned RITnet and U-Net models. We report results for four IPAD variants (L^1, L^2, L^1 LeGR and L^2 LeGR based) implemented with the combined filter-importance criterion $\psi(\cdot)$ at a fixed α, i.e., $\alpha = 0.5$, for consistency and fair comparisons, as no data-dependent optimization on the validation data is involved in this setting.

High-quality sclera segmentation: As shown in Fig. 10(a), in this simplest setting, all of the IPAD variants regardless of the targeted FLOP count lead to pruned segmentation models that outperform the original unpruned model. The exceptions here are the reference random and uniform pruning techniques. In the case of RITnet, the pruned models exhibit a particularly evident upward trend in performance when reducing the FLOP count to 75%, implying that the pruning of irrelevant filters does in fact initially bring a performance boost. However, after the first bundle of removed filters, the model's smaller and smaller size can no longer keep up with the problem's complexity and the performance slowly starts degrading. The trend in U-Net's case is less consistent, which also fits the above explanation, as U-Net is a much larger model. It starts with roughly 10× as many FLOPs as RITnet, and so even at 25% of the initial FLOPs, it is still more than twice as large as the unpruned RITnet. As such, the model initially exhibits the same performance boost achieved by the removal of irrelevant filters, but the decline after that is much less pronounced.

Limited-data in-the-wild sclera segmentation: In Fig. 10(b) we show the results on the more challenging SMD dataset, where a smaller number of training images with higher diversity is available for the experiments. We again observe that the pruned models quite consistently outperform the original unpruned model with most IPAD configurations. The upward trend on RITnet is still present, particularly at the 75% to 50% FLOP targets, while on U-Net this trend is less evident.

Cross-dataset sclera segmentation: The results of the cross-dataset experiments, where the models are trained on SMD and evaluated on SLD, are shown in Fig. 10(c). Because this cross-dataset segmentation problem is more challenging, we observe a different behavior of the pruned models. For most IPAD variants on the RITnet model, performance starts degrading slightly with any reduction in the models' FLOP counts. While some of the pruned models perform better at specific target FLOP percentages, the overall trend is towards weaker results. For the more heavily parameterized U-Net model the opposite can be observed. Here, the segmentation performance generally increases compared to the unpruned model for all FLOP targets, since even the smaller models are still large enough to generalize well to new and unseen ocular data.

Four-class ocular segmentation: The results of the four-class segmentation on the MOBIUS data, shown in Fig. 10(d), follow similar trends as in the previous experiments. The upward trend in the case of RITnet, particularly for the 75% to 50% FLOP targets, is clearly present, and most of the smaller models outperform the original unpruned version. With U-Net, we see a considerable boost in performance with the first reduction in complexity (at the 75% FLOP target) and then remains steady at smaller FLOP counts as well. Overall, the initial unpruned segmentation model once again gets significantly outperformed by all the smaller models, which agrees with our previous analysis.

4.5.3. Pruning criterion weighting

As shown in Sub Section 4.5.1, the evaluated pruning methods using our criterion relatively consistently beat the literature-standard weights-based criterion when using the optimal α determined on the validation data. In this section, we now study how the segmentation performance of the pruned models changes with different values of α used in the proposed combined filter-

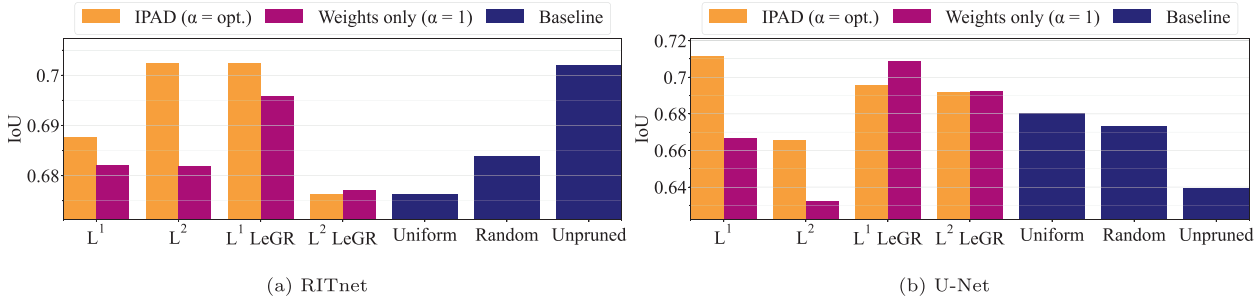

Fig. 8. Impact of the proposed filter-importance criterion $\psi(\cdot)$ on SLD. The segmentation performance in the *cross-dataset sclera segmentation* setting is reported for the pruning methods implemented with the proposed criterion vs. the standard weights-based criterion normally used in the literature.

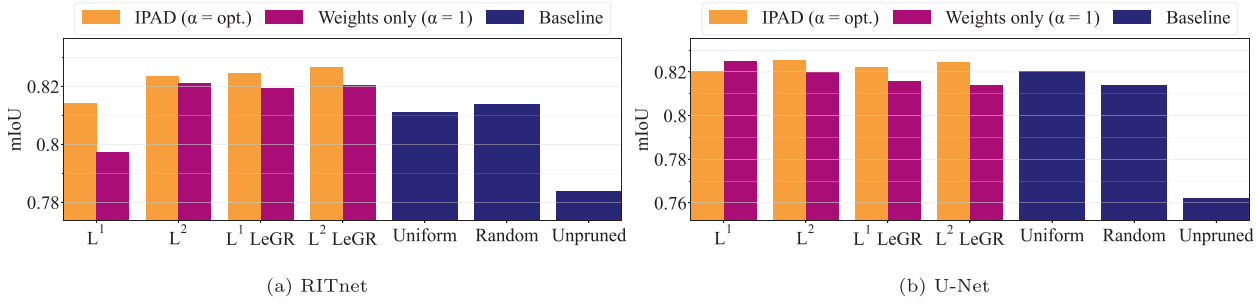

Fig. 9. Impact of the proposed filter-importance criterion $\psi(\cdot)$ on MOBIUS. The segmentation performance in the *four-class ocular segmentation* setting is reported for the pruning methods implemented with the proposed criterion vs. the standard weights-based criterion normally used in the literature.

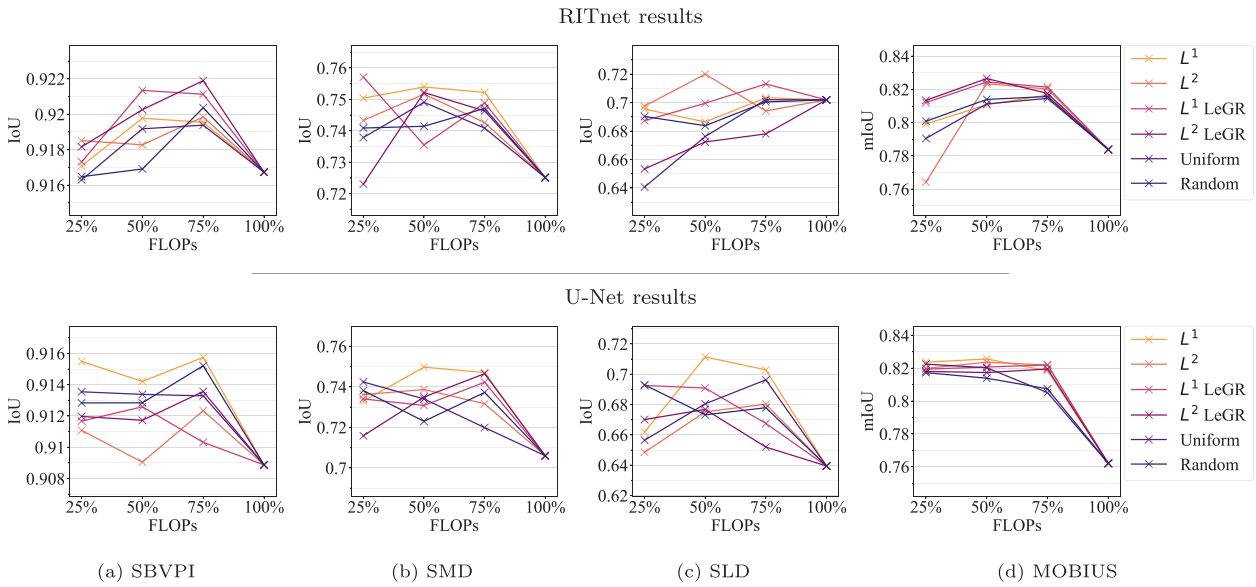

Fig. 10. Performance of the pruned models across different target FLOP counts. The right-most value in each graph is the performance of the unpruned model. The top row shows the results of RITnet (16 GFLOPs unpruned), while the bottom row contains the results of U-Net (160 GFLOPs unpruned).

importance criterion to determine how vital the proper selection of this balancing parameter is. We present the results for a fixed percentage of 50% FLOPs of the initial models for consistency. Additionally, we also report results for the uniform/random pruning approaches and the original unpruned model. It is important to note that these reference approaches do not rely on the value of α, and are therefore represented as horizontal lines in the presented graphs. Here, the left most point (for $\alpha = 0$) corresponds to using only the ADC criterion, the right most point (for $\alpha = 1$) to using only the weights-based criterion, whereas all other points represent possible operating points for the proposed combined filter-importance criterion. The experiments are again conducted for four different IPAD versions.

High-quality sclera segmentation: From the results in Fig. 11 (a) we can see that, while most IPAD variants have multiple values of α where they outperform the weights-based criterion (right-most point in the graphs), there also relatively consistently appear values of α that lead to somewhat worse performance. This implies that the selection of the correct α parameter is crucial for the pruning procedure and consequent segmentation performance. What is more, while the ranking of different values of α is quite similar for L^1 and L^1 LeGR or for L^2 and L^2 LeGR, this is not the case when comparing L^1-based methods to L^2-based methods. This observation suggests that the selection of the correct α depends mainly on the norms used in the calculation of the criterion.

Limited-data in-the-wild sclera segmentation: In Fig. 11(b) we can again see that the weights-based criterion (right-most value) outperforms some of the poorly chosen values of α in all the methods except in the case of L^1 pruning on RITnet. The L^1 methods and the L^2 ones once again follow similar trends, although the difference between the L^1 and L^2 methods is less pronounced than it was in the previous experiment. The selection of the α parameter is still observed to be an important factor for the higher level of performance exhibited by the segmentation models pruned based on the proposed combined filter-importance criterion.

Cross-dataset sclera segmentation: In this experiment, which focuses primarily on the models' ability to generalize to unseen data, the results in Fig. 11(c) tell a different story from the previous two experiments. This time, most of the combined-criterion based methods outperform their weight-based counterpart (right-most point) regardless of the choice of α. This observation implies that the proposed activation-based criterion, irrespective of how strongly it is weighted, critically contributes to the determination of filter importance and consequently leads to pruned models with better generalization capabilities. Additionally, note that in RITnet's case, all results corresponding to the weights-based criterion (the right-most points) are actually below the original unpruned model, as already discussed in Section 4.5.1. Conversely, several other variants of the evaluated pruning methods still lead to segmentation models that outperform the original model with the optimal choice of α. Interestingly, the choice of the IPAD variant seems to be particularly relevant for the selection of the α parameter in this experiment, especially with the U-Net segmentation model. As can be seen, L^1 LeGR and L^2 LeGR follow almost the same trend across different target FLOPs, and so do the L^1 and L^2 IPAD implementations. Given that the performance of the pruned models varies with respect to the values of α used in the implementation of the pruning method, a proper choice of the parameter again appears critical for good performance.

Four-class ocular segmentation: With the four-class results in Fig. 11(d), we observe more consistent performance across different choices of α than in the previous two experiments, with only L^1 IPAD pruning performing somewhat inconsistently. Here, the results with the best performing α are always better than the results achieved with the weight-based criterion only, whereas even the worst choice of α still leads to better performance than the purely weights-based criterion in 2 out 8 cases.

4.5.4. Ablation study

In the previous section, we explored the importance of the proper choice of the α parameter. In this ablation study, we study this specific aspect more explicitly. Specifically, we look at the performance differences if we *(i)* remove the weights-based criterion component (i.e. $\alpha = 0$), *(ii)* remove the activation-based component (i.e. $\alpha = 1$), *(iii)* remove the α selection process on validation data (as described in Section 4.5.1) and and use a predetermined fixed α instead (i.e. $\alpha = 0.5$). All the pruned models' results are reported at 50% of the FLOPs of the original models in Fig. 12.

- **No weights-based criterion**: The $\alpha = 0$ results in Fig. 12 show the effect of turning off the weights-based component of the proposed filter-importance criterion $\psi(\cdot)$ completely. In 25 of the 32 considered cases, the chosen combination of the two criterion components matches or outperforms the activation-based component alone. In 19 out of the 32 cases even the fixed $\alpha = 0.5$ combination outperforms the activation-based component alone, which is still more than half of all cases, although by a less significant margin.
- **No activation-based criterion**: The $\alpha = 1$ results of Fig. 12 show the impact of disabling the activation-based component of our criterion. In 27 of the 32 cases, the best α choice matches or outperforms the weights-based component alone, following a similar trend as observed in the previous experiment. In 21 of the 32 cases, the weights-based component alone is outperformed by the fixed $\alpha = 0.5$ combined criterion, again following a similar trend as observed in the previous experiment. Note that the weights-based component alone performed worse by 2 in both of these case counts than the activation-based component alone, again pointing to the superiority of our activation-based criterion.
- **No α selection**: At $\alpha = 0.5$ the results of Fig. 12 show how using a fixed α changes the results relative to determining the best α value on the validation data. In 27 of the 32 cases the best α choice matches or outperforms the fixed-value $\alpha = 0.5$ result, which is again the vast majority of the cases and points to the importance of optimizing the α parameter on some hold-out data.

Overall, the presented results suggest that, while our activation-based criterion component alone seems to perform slightly better than the standard weights-based criterion alone, the combination of the two is still the far superior choice. Additionally, the proper choice of α is shown to be crucial for the overall pruning process's success. However, even with a fixed-α combination of the two criteria, the pruned models in general exhibit a better performance than models pruned with either criterion alone.

Another important aspect of the pruning procedure is the type of convolutional layer the procedure is applied to. In general, 1×1 convolutions are intrinsically different from 3×3 (and other) convolutions, since they are typically used for channel mixing in dimensionality reduction and not spatial filtering. To this end, we next study the impact of: (*i*) completely removing the pruning of 1×1 convolutions from our pruning procedure, and (*ii*) pruning 1×1 convolutions but only applying our criterion to the 3×3 convolutions, while 1×1 convolutions in this case use the classic weights-based criterion only. Since U-Net only has 3×3 convolutions, we only report the results for this ablation on RITnet. We report the results in the bar graphs of Fig. 13 at 50% total FLOPs and the optimal α values selected on the validation data.

- **No channel pruning**: We note from Fig. 13 that removing 1×1 pruning quite significantly decreases the performance in general relative to the L^1 and L^2 based IPAD variants (denoted as Classic) as well as L^1 and L^2 LeGR IPAD methods. The only exception to this general trend is the cross-dataset experiment, where excluding the 1×1 convolutions from the pruning process still leads to the 2nd best result overall. In all other experiments, the removal of this pruning step is detrimental for the segmentation performance of the pruned models.
- **Classic weights-based criterion for channel pruning**: The milder version of still pruning 1×1 convolutions but only relying on the weights-based criterion for the pruning process (since our criterion was developed with spatial convolutions in mind) performs significantly better overall, outperforming all other methods in the two cases with very limited training data (SMD and SLD) and outperforms the procedure with 1×1 pruning fully removed in 6 out of 8 cases. However, with sufficient training data (SBVPI and MOBIUS), it is still outperformed consistently by the IPAD variants based on the classic methods as well as LeGR.

4.5.5. Qualitative comparison

Finally, we show a few examples of the model predictions before and after the pruning procedure in Fig. 14. The goal of these visualizations is to explore the impact of the pruning process on the behavior of the segmentation models. The results for all pruned

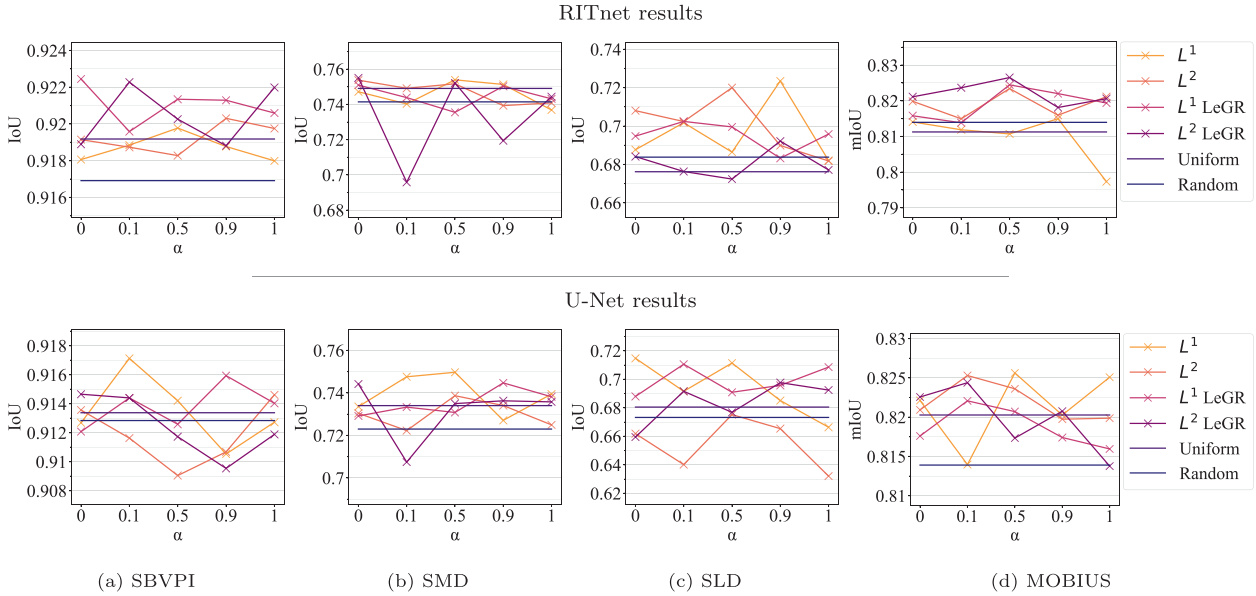

Fig. 11. Performance of the pruned models across different values of the α balancing parameter. The top row shows the results for RITnet, while the bottom row contains the results for U-Net. Points corresponding to $\alpha = 1$ correspond to the standard weights-based criterion, points corresponding to $\alpha = 0$ to the ADC criterion, and all the rest to different variants of the proposed criterion from Eq. (6).

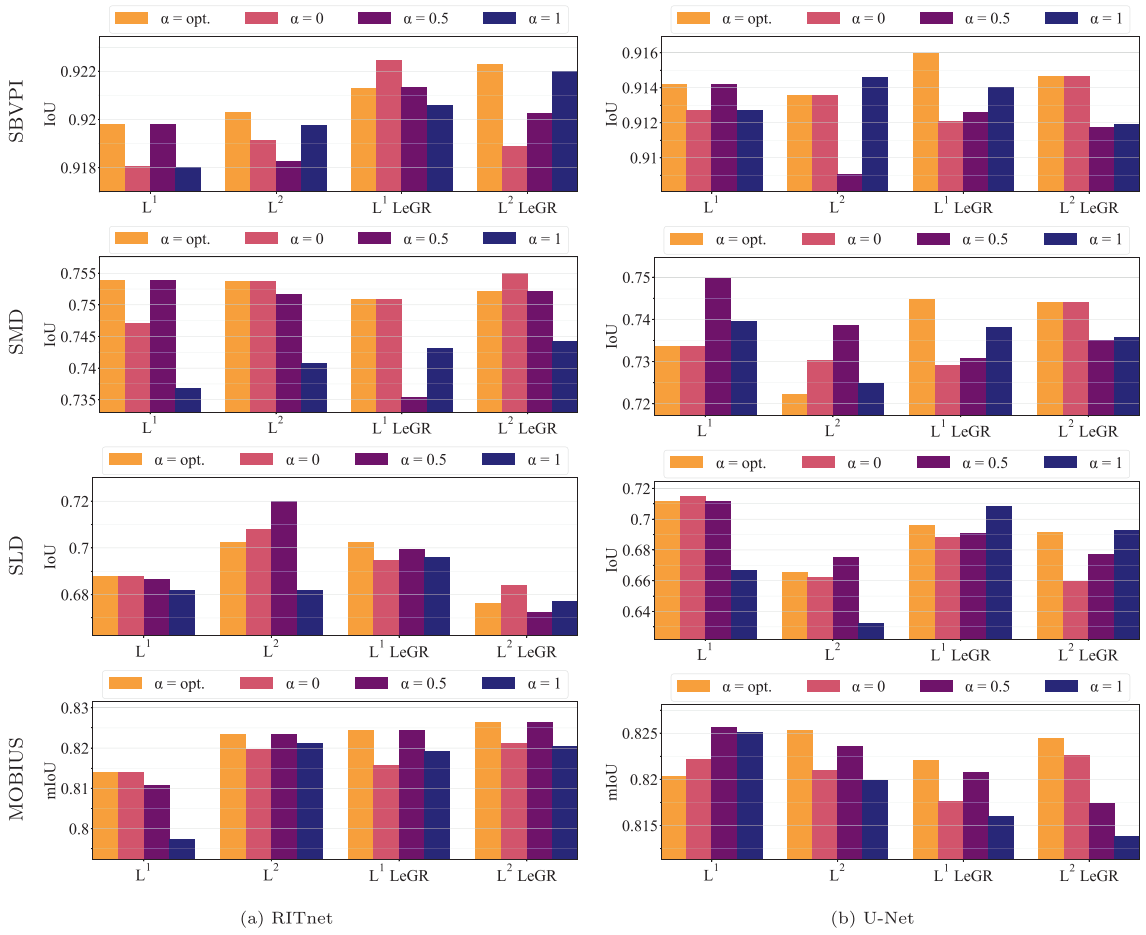

Fig. 12. Results of the ablation study. The graphs show segmentation performance differences when components of the proposed pruning process are selectively turned off. The results show the impact of removing the weights-based criterion component ($\alpha = 0$), removing the activation-based component ($\alpha = 1$), and removing the α selection process ($\alpha = 0.5$). The rows show the results on different datasets, in top-to-bottom order: SBVPI, SMD, SLD, MOBIUS.

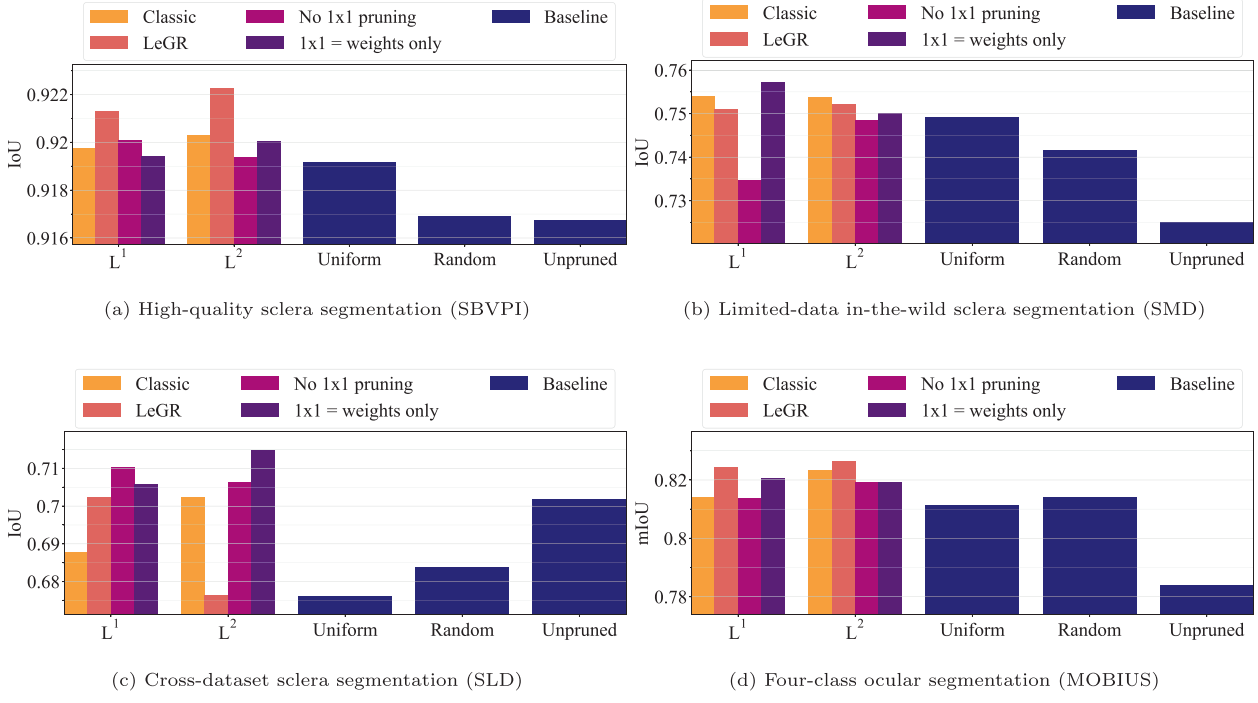

Fig. 13. Ablation study results *w.r.t.* **the filter-types pruned.** The results show the segmentation performance of pruned RITnet models with different variants of IPAD when the pruning of 1×1 filters is completely disabled or when the weights-based criterion is used to prune the 1×1 filters.

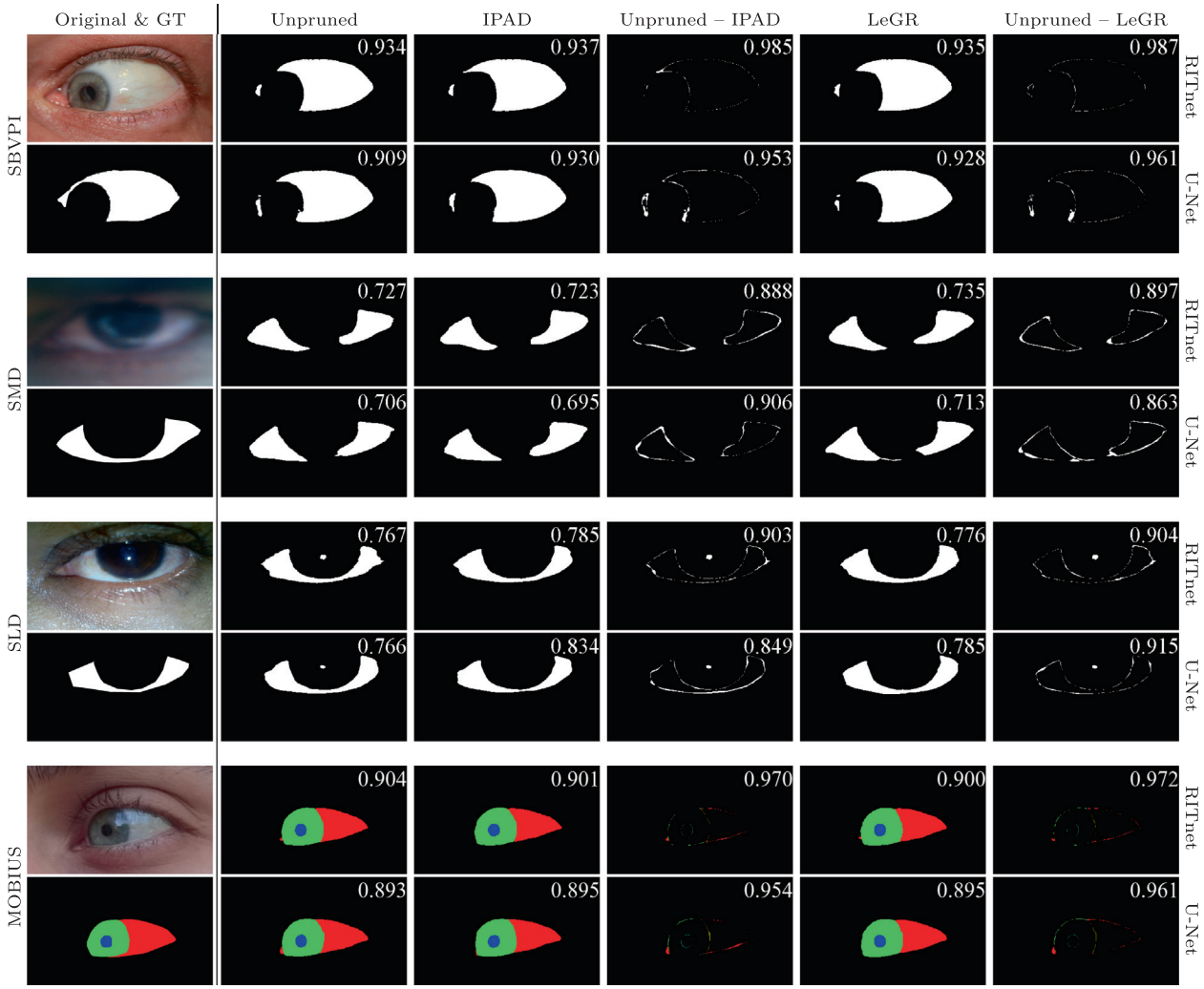

Fig. 14. Qualitative comparison of the model predictions. The four blocks of images (along the rows) show visual segmentation results on the four experimental datasets. The columns show the predictions of the models and the differences between the unpruned and pruned models for the L^2 IPAD variant and the original L^2 LeGR method. The difference images (4th and 6th column) show the difference in the predictions between the pruned and unpruned models. The predictions are overlayed with their *IoU* scores and the difference images are overlayed with the *IoU* between the pruned and unpruned models' predictions.

models are shown at 50% the FLOPs of the initial unpruned model. Note that the pruned model predictions stay close to the original unpruned model. Even with poor segmentation results, such as the ones presented for the SMD and SLD datasets, the *IoU* between the pruned models' predictions and the original unpruned model's prediction remains high. This consistency of predictions is precisely the goal of the pruning procedure. Also note how the false positive in the SLD block, that appears with both RITnet and U-Net due to the specular reflection in the original image, is removed by either pruning procedure for both segmentation models, showing the advantage of simplifying the models through pruning, as discussed throughout Section 4.5.2.

5. Conclusion

In this paper, we presented a novel criterion for determining filter importance in convolutional neural networks (CNNs) in the process of filter pruning and designed an iterative pruning procedure around this novel criterion. We evaluated the proposed criterion in four distinct problem settings: high-quality-data sclera segmentation, limited-data mobile sclera segmentation, cross-dataset sclera segmentation, and 4-class mobile ocular segmentation. We tested the proposed approach with two deep learning models of significantly different initial sizes (RITnet at 16 initial GFLOPs and U-Net at 160) with highly encouraging results.

Our criterion consistently outperformed the classic L^1 and L^2 weights-based criteria from the literature, as well as the uniform and random pruning baselines, and in most cases even the original unpruned model. Despite the small absolute differences in performance, the increase in performance is consistent and significant – for a full statistical analysis of the significance of this increase, we refer the reader to the Appendix. This implies that the proposed criterion better determines, which filters are of low importance to the overall segmentation model and can therefore be pruned away.

As part of our future work we plan to explore the use of the proposed criterion in different tasks that require a measure of filter importance, such as CNN visualization and interpretability, selective knowledge distillation, transfer learning, and others. Additionally, because our method measures the amount of new information of a filter through deviations from the mean activations, linear shifts in the activation maps may produce high importance scores, even though this may not necessarily imply a high degree of new

information. Thus, we also plan to explore extensions of our criterion based on higher-order statistics in the scope of our future research activities to address this limitation.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: [The research was partially funded by the Slovenian Research Agency (ARRS).]

Acknowledgements

Supported by the ARRS Research Programme P2-0250(B) "Metrology and Biometric Systems", the ARRS Research Programme P2-0214 "Computer Vision", the ARRS Young Researcher Programme, and NVIDIA's "Academic Hardware Grants" Programme, which provided two of the graphic cards used in our research.

Appendix A

To demonstrate the reliability of the performance increase our novel criterion brings to the pruning process, we perform a statistical analysis of the obtained results in this appendix. Table A1 shows all the numerical results from the main experiments discussed in Section 4.5. If we assume a random distribution of the results (i.e. no impact from our criterion), we would expect the best outcome to be in the right-most column in one fifth of the experiments. Since there are 96 rows with α-based results in total, we would expect roughly $\frac{1}{5} \cdot 96 \approxeq 19$ of them to have the best

Table A1
Full $(m)IoU$ results of the pruning experiments on the 4 evaluation datasets. The best result in each row is presented in bold, while the best result of each block is coloured red.

(a) MOBIUS

Model	FLOPs	Method	α=0	α=0.1	α=0.5	α=0.9	α=1
RITnet	25%	L1	80.81%	**80.98%**	79.92%	80.58%	80.19%
		L1 LeGR	80.55%	80.84%	81.20%	**81.93%**	80.95%
		L2	76.36%	77.41%	76.42%	**78.63%**	76.08%
		L2 LeGR	**81.81%**	81.63%	81.34%	79.23%	78.69%
		Random			80.08%		
		Uniform			79.05%		
	50%	L1	81.41%	81.19%	81.07%	**81.50%**	79.73%
		L1 LeGR	81.58%	81.38%	**82.45%**	82.20%	81.94%
		L2	81.98%	81.49%	**82.34%**	81.60%	82.11%
		L2 LeGR	82.11%	82.37%	**82.65%**	81.81%	82.06%
		Random			81.40%		
		Uniform			81.13%		
	75%	L1	81.41%	81.77%	81.66%	81.48%	**82.01%**
		L1 LeGR	80.81%	**82.29%**	82.14%	81.86%	82.29%
		L2	**82.26%**	81.88%	82.02%	82.22%	81.93%
		L2 LeGR	**82.30%**	82.00%	81.76%	82.03%	81.85%
		Random			81.57%		
		Uniform			81.46%		
U-Net	25%	L1	**82.61%**	82.55%	82.36%	82.56%	82.19%
		L1 LeGR	81.85%	81.88%	**81.94%**	81.31%	81.24%
		L2	82.18%	82.09%	81.99%	**82.29%**	82.12%
		L2 LeGR	81.50%	80.13%	**81.80%**	80.80%	81.67%
		Random			81.73%		
		Uniform			82.25%		
	50%	L1	82.21%	81.40%	**82.56%**	82.03%	82.51%
		L1 LeGR	81.76%	**82.21%**	82.07%	81.74%	81.60%
		L2	82.09%	**82.53%**	82.36%	81.97%	81.99%
		L2 LeGR	82.26%	**82.44%**	81.73%	82.08%	81.38%
		Random			81.39%		
		Uniform			82.03%		
	75%	L1	**82.34%**	82.23%	81.85%	82.08%	82.18%
		L1 LeGR	81.87%	81.99%	**82.20%**	81.55%	81.45%
		L2	81.93%	81.73%	82.20%	**82.37%**	82.04%
		L2 LeGR	81.84%	**82.00%**	81.95%	81.58%	81.87%
		Random			80.74%		
		Uniform			80.51%		

(b) SBVPI

Model	FLOPs	Method	α=0	α=0.1	α=0.5	α=0.9	α=1
RITnet	25%	L1	91.62%	91.71%	91.71%	**91.80%**	91.74%
		L1 LeGR	91.59%	91.57%	91.73%	**92.02%**	91.95%
		L2	91.72%	91.71%	**91.85%**	91.76%	91.77%
		L2 LeGR	92.05%	91.98%	91.82%	**92.08%**	91.87%
		Random			91.65%		
		Uniform			91.63%		
	50%	L1	91.81%	91.88%	**91.98%**	91.88%	91.80%
		L1 LeGR	**92.24%**	91.96%	92.13%	92.13%	92.06%
		L2	91.91%	91.87%	91.83%	**92.03%**	91.98%
		L2 LeGR	91.89%	**92.23%**	92.03%	91.88%	92.20%
		Random			91.69%		
		Uniform			91.92%		
	75%	L1	**92.08%**	92.05%	91.95%	92.02%	91.98%
		L1 LeGR	92.13%	92.19%	92.11%	**92.23%**	92.20%
		L2	**92.11%**	92.01%	91.98%	91.77%	91.86%
		L2 LeGR	92.09%	**92.19%**	92.19%	92.14%	92.19%
		Random			92.04%		
		Uniform			91.94%		
U-Net	25%	L1	91.44%	91.54%	**91.55%**	91.02%	91.13%
		L1 LeGR	91.35%	91.31%	91.17%	**91.37%**	91.11%
		L2	91.22%	90.94%	91.11%	**91.34%**	91.16%
		L2 LeGR	91.12%	91.32%	91.20%	**91.37%**	91.16%
		Random			91.28%		
		Uniform			91.35%		
	50%	L1	91.27%	**91.71%**	91.42%	91.05%	91.27%
		L1 LeGR	91.21%	91.44%	91.26%	**91.59%**	91.40%
		L2	91.36%	91.16%	90.90%	91.07%	**91.46%**
		L2 LeGR	**91.47%**	91.44%	91.17%	90.95%	91.19%
		Random			91.28%		
		Uniform			91.34%		
	75%	L1	91.38%	91.56%	**91.57%**	91.43%	91.32%
		L1 LeGR	91.20%	91.22%	91.03%	91.25%	**91.27%**
		L2	**91.40%**	91.18%	91.23%	91.01%	91.27%
		L2 LeGR	91.01%	91.01%	**91.36%**	91.26%	91.10%
		Random			91.52%		
		Uniform			91.33%		

(c) SMD

Model	FLOPs	Method	α=0	α=0.1	α=0.5	α=0.9	α=1
RITnet	25%	L1	74.95%	74.08%	**75.04%**	74.55%	74.63%
		L1 LeGR	72.91%	70.38%	**75.71%**	74.07%	75.67%
		L2	72.75%	**74.92%**	74.33%	74.05%	72.91%
		L2 LeGR	70.53%	72.61%	72.30%	**74.21%**	73.79%
		Random			74.09%		
		Uniform			73.78%		
	50%	L1	74.72%	74.03%	**75.40%**	75.13%	73.68%
		L1 LeGR	**75.09%**	74.39%	73.55%	75.03%	74.32%
		L2	**75.38%**	74.91%	75.17%	73.93%	74.09%
		L2 LeGR	**75.51%**	69.58%	75.22%	71.94%	74.43%
		Random			74.14%		
		Uniform			74.91%		
	75%	L1	75.90%	75.11%	75.22%	**76.80%**	76.13%
		L1 LeGR	73.84%	75.11%	74.89%	**75.20%**	74.91%
		L2	**75.86%**	75.24%	74.26%	74.46%	74.95%
		L2 LeGR	73.89%	73.39%	74.63%	**74.65%**	73.60%
		Random			74.71%		
		Uniform			74.09%		
U-Net	25%	L1	72.66%	72.71%	**73.29%**	73.19%	72.60%
		L1 LeGR	74.67%	**75.37%**	73.40%	73.81%	73.50%
		L2	72.85%	72.73%	**73.61%**	72.45%	73.09%
		L2 LeGR	**76.08%**	72.11%	71.59%	72.71%	72.35%
		Random			73.81%		
		Uniform			74.24%		
	50%	L1	73.35%	74.76%	**74.97%**	72.71%	73.95%
		L1 LeGR	72.91%	73.33%	73.08%	**74.47%**	73.82%
		L2	73.04%	72.23%	**73.87%**	73.41%	72.50%
		L2 LeGR	**74.42%**	70.74%	73.50%	73.63%	73.58%
		Random			72.30%		
		Uniform			73.40%		
	75%	L1	72.74%	73.94%	**74.70%**	73.06%	72.98%
		L1 LeGR	**74.30%**	73.00%	74.23%	73.41%	72.79%
		L2	71.86%	**73.19%**	73.17%	72.14%	73.06%
		L2 LeGR	72.63%	72.78%	**74.65%**	72.82%	71.91%
		Random			73.72%		
		Uniform			71.99%		

(d) SLD

Model	FLOPs	Method	α=0	α=0.1	α=0.5	α=0.9	α=1
RITnet	25%	L1	**70.42%**	67.78%	69.54%	61.94%	67.30%
		L1 LeGR	64.58%	69.94%	68.73%	**70.49%**	68.82%
		L2	**70.95%**	66.74%	69.76%	61.34%	70.82%
		L2 LeGR	65.22%	60.98%	65.34%	**69.74%**	69.24%
		Random			69.03%		
		Uniform			64.07%		
	50%	L1	68.77%	70.17%	68.64%	**72.35%**	68.20%
		L1 LeGR	69.47%	**70.24%**	69.95%	68.32%	69.58%
		L2	70.80%	70.23%	**72.00%**	68.99%	68.18%
		L2 LeGR	68.41%	67.63%	67.24%	**69.21%**	67.71%
		Random			68.38%		
		Uniform			67.62%		
	75%	L1	67.75%	69.95%	70.34%	**71.71%**	69.05%
		L1 LeGR	68.75%	68.57%	**71.31%**	67.01%	68.57%
		L2	69.90%	68.88%	69.41%	**71.54%**	66.77%
		L2 LeGR	**69.56%**	65.09%	67.80%	65.34%	66.51%
		Random			70.05%		
		Uniform			70.21%		
U-Net	25%	L1	**70.11%**	69.72%	66.19%	67.20%	67.35%
		L1 LeGR	**70.19%**	63.71%	69.26%	69.01%	66.45%
		L2	**65.75%**	63.36%	64.86%	62.03%	65.67%
		L2 LeGR	**70.95%**	68.69%	67.01%	70.83%	68.52%
		Random			69.30%		
		Uniform			65.65%		
	50%	L1	**71.47%**	69.16%	71.14%	68.51%	66.64%
		L1 LeGR	68.80%	**71.06%**	69.09%	69.57%	70.85%
		L2	66.18%	64.01%	**67.52%**	66.54%	63.21%
		L2 LeGR	65.95%	69.16%	67.68%	**69.76%**	69.24%
		Random			67.32%		
		Uniform			68.05%		
	75%	L1	**72.22%**	69.61%	70.29%	64.47%	66.06%
		L1 LeGR	66.06%	**70.54%**	66.75%	69.22%	69.92%
		L2	66.33%	64.79%	**68.02%**	66.20%	66.09%
		L2 LeGR	**69.27%**	68.09%	65.20%	69.20%	67.65%
		Random			67.79%		
		Uniform			69.63%		

result in the right-most column. However, as we see in the table, that is only the case for 3 of the rows. As also illustrated in Fig. A1, the probability of there being at most 3 best results in the right-most column at random is roughly 10^{-6}.

A similar analysis can be conducted for the blocks of Table A1. We notice that none of the block-best results appear in either the right-most column or the uniform/random pruning rows. There are a total of 24 blocks with 22 distinct numerical results each. The probability of the best result of a block being in the right-most column or the uniform/random row in the case of random results is

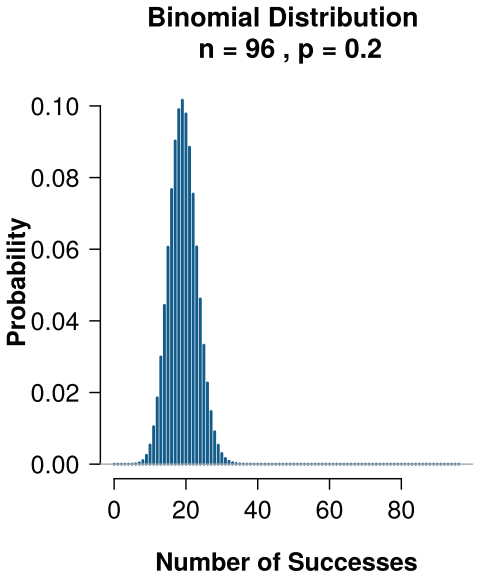

Fig. A1. A histogram of the binomial distribution $X \sim Bin(96, \frac{1}{5})$, which corresponds to 96 trials with a success implying the highest achieved score was in the right-most column in Table A1 (corresponding to $\alpha = 1$, i.e. not using our criterion). As we can see, the probability of $\leqslant 3$ successes in the case of random results (i.e. $P(x) = \frac{1}{5}$.) is negligible.

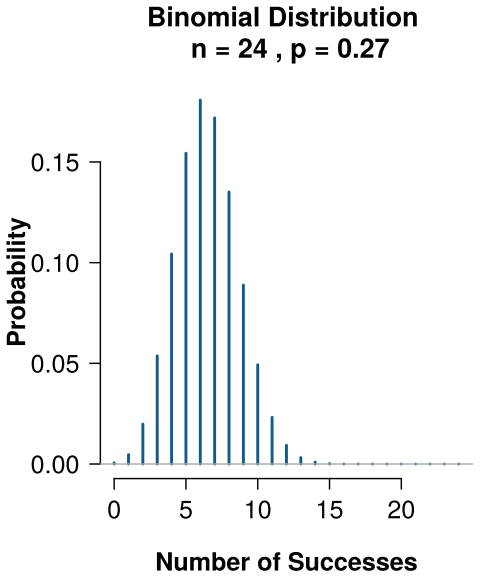

Fig. A2. A histogram of the binomial distribution $X \sim Bin(24, \frac{6}{22})$, corresponding to 24 trials with a success implying the highest achieved score was in the right-most column or in the uniform/random row in Table A1 (which corresponds to $\alpha = 1$, i.e. not using our criterion). As we can see, the probability of no successes in the case of random results (i.e. $P(x) = \frac{6}{22}$.) is very small.

$\frac{6}{22} \approx 0.27$. The distribution is shown in Fig. A2 and the probability of 0 best results in the right-most column or the uniform/random row at random is $(\frac{16}{22})^{24} \approx 5 \cdot 10^{-4}$.

With both these observations, we can conclude with reasonable statistical certainty that our criterion indeed provides a reliable and consistent performance boost.

References

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) 39 (12), 2481–2495.

Blake, J., Maguire, L.P., McGinnity, T., Roche, B., McDaid, L., 1998. The implementation of fuzzy systems, neural networks and fuzzy neural networks using FPGAs. Informat. Sci. 112 (1–4), 151–168.

Boutros, F., Damer, N., Kirchbuchner, F., Kuijper, A., 2019. Eye-mms: Miniature multi-scale segmentation network of key eye-regions in embedded applications. In: IEEE/CVF International Conference on Computer Vision Workshops (ICCVW).

Brigato, L., Iocchi, L., 2020. A close look at deep learning with small data. In: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, pp. 2490–2497.

Chang, Y.-H., Lee, G.G., Chen, S.-Y., 2022. Deep learning acceleration design based on low rank approximation. In: 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE, pp. 1304–1307.

Chaudhary, A.K., Kothari, R., Acharya, M., Dangi, S., Nair, N., Bailey, R., Kanan, C., Diaz, G., Pelz, J.B., 2019. RITnet: Real-time Semantic Segmentation of the Eye for Gaze Tracking. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), IEEE, pp. 3698–3702.

Chen, G., Choi, W., Yu, X., Han, T., Chandraker, M., 2017. Learning efficient object detection models with knowledge distillation. In: Advances in Neural Information Processing Systems, pp. 742–751.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2018. DeepLab: Semantic Image Segmentation With Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. 40 (4), 834–848.

Cheng, Y., Wang, D., Zhou, P., Zhang, T., 2018. Model compression and acceleration for deep neural networks: The principles, progress, and challenges. IEEE Signal Process. Mag. 35 (1), 126–136.

Chin, T.-W., Ding, R., Zhang, C., Marculescu, D., 2020. Towards Efficient Model Compression via Learned Global Ranking. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Choudhary, T., Mishra, V., Goswami, A., Sarangapani, J., 2020. A comprehensive survey on model compression and acceleration. Artif. Intell. Rev., 1–43

Das, A., 2017. Towards Multi-modal Sclera and Iris Biometric Recognition with Adaptive Liveness Detection Ph.D. thesis. Griffith University.

Das, A., Pal, U., Blumenstein, M., Ballester, M.A.F., 2013. Sclera recognition-a survey. In: 2013 2nd IAPR Asian Conference on Pattern Recognition, pp. 917–921.

Das, A., Pal, U., Ferrer, M.A., Blumenstein, M., 2015. SSBC 2015: Sclera Segmentation Benchmarking Competition. In: Conference on Biometrics: Theory, Applications, and Systems (BTAS), pp. 742–747.

Das, A., Pal, U., Ferrer-Ballester, M.A., Blumenstein, M., 2016. SSRBC 2016: Sclera Segmentation and Recognition Benchmarking Competition. In: International Conference on Biometrics (ICB), pp. 1–6.

Das, A., Pal, U., Ferrer, M.A., Blumenstein, M., Štepec, D., Rot, P., Emeršič, Z., Peer, P., Štruc, V., Kumar, S., 2017. SSERBC 2017: Sclera segmentation and eye recognition benchmarking competition. In: International Joint Conference on Biometrics (IJCB), pp. 742–747.

Das, A., Pal, U., Ferrer, M.A., Blumenstein, M., Štepec, D., Rot, P., Peer, P., Štruc, V., 2018. SSBC 2018: Sclera segmentation benchmarking competition. In: International Conference on Biometrics (ICB), pp. 303–308.

Das, A., Pal, U., Blumenstein, M., Wang, C., He, Y., Zhu, Y., Sun, Z., 2019. Sclera Segmentation benchmarking competition in cross-resolution environment. In: IAPR International Conference on Biometrics. IEEE.

Das, S., De Ghosh, I., Chattopadhyay, A., 2022. Sclera biometrics in restricted and unrestricted environment with cross dataset evaluation. Displays 74, 102257.

Derakhshani, R., Ross, A., 2007. A texture-based neural network classifier for biometric identification using ocular surface vasculature. In: International Joint Conference on Neural Networks 2007 (IJCNN 2007), IEEE, pp. 2982–2987.

Dimauro, G., Camporeale, M.G., Dipalma, A., Guarini, A., Maglietta, R., 2023. Anaemia detection based on sclera and blood vessel colour estimation. Biomed. Signal Process. Control 81, 104489.

Dupuis, E., Novo, D., O'Connor, I., Bosio, A., 2021. CNN weight sharing based on a fast accuracy estimation metric. Microelectron. Reliab. 122, 114148.

Dupuis, E., Novo, D., O'Connor, I., Bosio, A., 2022. A heuristic exploration of retraining-free weight-sharing for CNN compression. In: 2022 27th Asia and South Pacific Design Automation Conference (ASP-DAC), IEEE, pp. 134–139.

Garbin, S.J., Shen, Y., Schuetz, I., Cavin, R., Hughes, G., Talathi, S.S., 2019. OpenEDS: Open Eye Dataset, arXiv preprint arXiv:1905.03702.

Gong, Y., Khurana, S., Rouditchenko, A., Glass, J., 2002. Cmkd: Cnn/transformer-based cross-model knowledge distillation for audio classification, arXiv preprint arXiv:2203.06760.
Gysel, P., Motamedi, M., Ghiasi, S., 2016. Hardware-oriented approximation of convolutional neural networks.
Han, S., Mao, H., Dally, W.J., 2016. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In: International Conference on Learning Representations (ICLR).
He, Y., Zhang, X., Sun, J., 2017. Channel pruning for accelerating very deep neural networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1389–1397.
He, Y., Kang, G., Dong, X., Fu, Y., Yang, Y., 2018. Soft filter pruning for accelerating deep convolutional neural networks. In: International Joint Conference on Artificial Intelligence (IJCAI).
Hinton, G., Vinyals, O., Dean, J., 2014. Distilling the knowledge in a neural network. In: Neural Information Processing Systems (NeurIPS) Deep Learning Workshop.
Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4700–4708.
Huang, Q., Zhou, K., You, S., Neumann, U., 2018. Learning to prune filters in convolutional neural networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, pp. 709–718.
Hu, H., Peng, R., Tai, Y.-W., Tang, C.-K., 2016. Network trimming: A data-driven neuron pruning approach towards efficient deep architectures, arXiv preprint arXiv:1607.03250.
Hu, Y., Huang, T., Run, R., Yin, L., Li, G., Xie, X., 2022. PPBAM: A preprocessing-based power-efficient approximate multiplier design for CNN. In: 2022 IEEE International Conference on Integrated Circuits, Technologies and Applications (ICTA), IEEE, pp. 166–167.
IEEE Standard for Floating-Point Arithmetic, 2008. IEEE Std 754-2008, pp. 1–70. https://doi.org/10.1109/IEEESTD.2008.4610935.
Jaderberg, M., Vedaldi, A., Zisserman, A., 2014. Speeding up convolutional neural networks with low rank expansions. In: Proceedings of the British Machine Vision Conference. BMVA Press.
Kim, M.S., Del Barrio, A.A., Oliveira, L.T., Hermida, R., Bagherzadeh, N., 2018. Efficient mitchell's approximate log multipliers for convolutional neural networks. IEEE Trans. Comput. 68 (5), 660–675.
Kozyrskiy, N., Phan, A.-H., 2020. CNN acceleration by low-rank approximation with quantized factors, arXiv preprint arXiv:2006.08878.
LeCun, Y., Denker, J.S., Solla, S.A., 1990. Optimal brain damage. In: Advances in Neural Information Processing Systems, pp. 598–605.
Liang, T., Glossner, J., Wang, L., Shi, S., Zhang, X., 2021. Pruning and quantization for deep neural network acceleration: A survey. Neurocomputing 461, 370–403.
Li, H., Kadav, A., Durdanovic, I., Samet, H., Graf, H.P., 2017. Pruning filters for efficient convnets. In: International Conference on Learning Representations (ICLR).
Li, Y., Adamczewski, K., Li, W., Gu, S., Timofte, R., Van Gool, L., 2022. Revisiting Random Channel Pruning for Neural Network Compression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 191–201.
Lin, G., Milan, A., Shen, C., Reid, I., 2017. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
Liu, Z., Sun, M., Zhou, T., Huang, G., Darrell, T., 2018. Rethinking the value of network pruning. In: International Conference on Learning Representations.
Liu, Y., Chen, K., Liu, C., Qin, Z., Luo, Z., Wang, J., 2019. Structured Knowledge Distillation for Semantic Segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2604–2613.
Lotrič, U., Bulić, P., 2012. Applicability of approximate multipliers in hardware neural networks. Neurocomputing 96, 57–65.
Lozej, J., Meden, B., Štruc, V., Peer, P., 2018. End-to-end iris segmentation using U-Net. In: 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI), IEEE, pp. 1–6.
Luo, J.-H., Wu, J., 2017. An entropy-based pruning method for cnn compression, arXiv preprint arXiv:1706.05791.
Luo, P., Zhu, Z., Liu, Z., Wang, X., Tang, X., 2016. Face model compression by distilling knowledge from neurons. In: Thirtieth AAAI Conference on Artificial Intelligence.
Luo, J.-H., Wu, J., Lin, W., 2017. Thinet: A filter level pruning method for deep neural network compression. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5058–5066.
Lv, W., Song, Y., Fu, R., Lin, X., Su, Y., Jin, X., Yang, H., Shan, X., Du, W., Huang, Q., et al., 2022. Deep learning algorithm for automated detection of polycystic ovary syndrome using scleral images. Front. Endocrinol. 12, 1869.
Masadeh, M., Hasan, O., Tahar, S., 2018. Comparative study of approximate multipliers, in. In: Proceedings of the 2018 on Great Lakes Symposium on VLSI. ACM, pp. 415–418.
Mei, S., Chen, X., Zhang, Y., Li, J., Plaza, A., 2021. Accelerating convolutional neural network-based hyperspectral image classification by step activation quantization. IEEE Trans. Geosci. Remote Sens. 60, 1–12.
Nevarez, Y., Beering, A., Najafi, A., Najafi, A., Yu, W., Chen, Y., Krieger, K.-L., Garcia-Ortiz, A., 2023. CNN Sensor Analytics with Hybrid-Float6 Quantization on Low-Power Embedded FPGAs. IEEE Access.
Neyshabur, B., Tomioka, R., Srebro, N., 2015. In search of the real inductive bias: on the role of implicit regularization in deep learning. In: International Conference on Learning Representations (ICLR) Workshop.
Nigam, I., Vatsa, M., Singh, R., 2015. Ocular biometrics: A survey of modalities and fusion approaches. Informat. Fusion 26, 1–35.
Novak, R., Bahri, Y., Abolafia, D.A., Pennington, J., Sohl-Dickstein, J., 2018. Sensitivity and generalization in neural networks: an empirical study. In: International Conference on Learning Representations.
Perry, J., Fernandez, A., 2019. Minenet: A dilated cnn for semantic segmentation of eye features. In: IEEE/CVF International Conference on Computer Vision Workshops (ICCVW).
Polyak, A., Wolf, L., 2015. Channel-level acceleration of deep face representations. IEEE Access 3, 2163–2175.
Riccio, D., Brancati, N., Frucci, M., Gragnaniello, D., 2017. An unsupervised approach for eye sclera segmentation. In: Iberoamerican Congress on Pattern Recognition. Springer, pp. 550–557.
Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y., 2015. Fitnets: Hints for thin deep nets. In: International Conference on Learning Representations (ICLR).
Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention (MICCAI): 18th International Conference, Proceedings, Part III. Springer International Publishing, Cham, pp. 234–241.
Rot, P., Emeršič, v., Štruc, V., Peer, P., 2018. Deep Multi-class Eye Segmentation for Ocular Biometrics. In: IEEE International Work Conference on Bioinspired Intelligence (IWOBI), pp. 1–8. https://doi.org/10.1109/IWOBI.2018.8464133.
Rot, P., Vitek, M., Grm, K., Emeršič, V., Peer, P., Štruc, V., 2020. Deep sclera segmentation and recognition. In: Uhl, A., Busch, C., Marcel, S., Veldhuis, R.N.J. (Eds.), Handbook of Vascular Biometrics (HVB). Springer, pp. 395–432. https://doi.org/10.1007/978-3-030-27731-4_13.
Schmid, F., Koutini, K., Widmer, G., 2023. Efficient large-scale audio tagging via transformer-to-cnn knowledge distillation. In: ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 1–5.
Shang, H., Wu, J.-L., Hong, W., Qian, C., 2022. Neural Network Pruning by Cooperative Coevolution, in. In: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI).
Stanton, S., Izmailov, P., Kirichenko, P., Alemi, A.A., Wilson, A.G., 2021. Does knowledge distillation really work? Adv. Neural Informat. Process. Syst. 34, 6906–6919.
Tai, C., Xiao, T., Zhang, Y., Wang, X., Weinan, E., 2016. Convolutional neural networks with low-rank regularization. In: International Conference on Learning Representations (ICLR).
Vitek, M., Rot, P., Štruc, V., Peer, P., 2020a. A Comprehensive Investigation into Sclera Biometrics: A Novel Dataset and Performance Study. Neural Comput. Appl. (NCAA), 17941–17955. https://doi.org/10.1007/s00521-020-04782-1.
Vitek, M., Das, A., Pourcenoux, Y., Missler, A., Paumier, C., Das, S., De Ghosh, I., Lucio, D.R., Zanlorensi Jr., L.A., Menotti, D., Boutros, F., Damer, N., Grebe, J.H., Kuijper, A., Hu, J., He, Y., Wang, C., Liu, H., Wang, Y., Sun, Z., Osorio-Roig, D., Rathgeb, C., Busch, C., Tapia Farias, J., Valenzuela, A., Zampoukis, G., Tsochatzidis, L., Pratikakis, I., Nathan, S., Suganya, R., Mehta, V., Dhall, A., Raja, K., Gupta, G., Khiarak, J.N., Akbari-Shahper, M., Jaryani, F., Asgari-Chenaghlu, M., Vyas, R., Dakshit, S., Dakshit, S., Peer, P., Pal, U., Štruc, V., 2020b. SSBC 2020: Sclera segmentation benchmarking competition in the mobile environment. In: IEEE International Joint Conference on Biometrics (IJCB), pp. 1–10, https://doi.org/10.1109/IJCB48548.2020.9304881.
Vitek, M., Das, A., Lucio, D.R., Zanlorensi, L.A., Menotti, D., Khiarak, J.N., Shahpar, M. A., Asgari-Chenaghlu, M., Jaryani, F., Tapia, J.E., Valenzuela, A., Wang, C., Wang, Y., He, Z., Sun, Z., Boutros, F., Damer, N., Grebe, J.H., Kuijper, A., Raja, K., Gupta, G., Zampoukis, G., Tsochatzidis, L., Pratikakis, I., Aruna Kumar, S., Harish, B., Pal, U., Peer, P., Štruc, V., 2023. Exploring bias in sclera segmentation models: a group evaluation approach. IEEE Trans. Informat. Forens. Sec. (TIFS) 18, 190–205. https://doi.org/10.1109/TIFS.2022.3216468.
Wang, C., He, Y., Liu, Y., He, Z., He, R., Sun, Z., 2019. ScleraSegNet: an improved U-net model with attention for accurate sclera segmentation. In: IAPR International Conference on Biometrics, vol. 1.
Wang, Y., Wang, J., Guo, P., 2022. Eye-UNet: a UNet-based network with attention mechanism for low-quality human eye image segmentation. Signal, Image Video Process., 1–7
Wu, R., Zhang, F., Zheng, Z., Du, X., Shen, X., 2021. Exploring deep reuse in winograd CNN inference. In: Proceedings of the 26th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, pp. 483–484.
Wu, R., Zhang, F., Guan, J., Zheng, Z., Du, X., Shen, X., 2022. Drew: Efficient winograd cnn inference with deep reuse. In: Proceedings of the ACM Web Conference, 2022, pp. 1807–1816.
Yang, T., Liao, Y., Shi, J., Liang, Y., Jing, N., Jiang, L., 2020. A Winograd-based CNN accelerator with a fine-grained regular sparsity pattern. In: Proceedings of the 30th International Conference on Field-Programmable Logic and Applications (FPL), IEEE, pp. 254–261.
Yi, L., Su, H., Guo, X., Guibas, L.J., 2017. Syncspeccnn: Synchronized spectral cnn for 3d shape segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2282–2290.
Yim, J., Joo, D., Bae, J., Kim, J., 2017. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4133–4141.
Young, S.I., Zhe, W., Taubman, D., Girod, B., 2021. Transform quantization for CNN compression. IEEE Trans. Pattern Anal. Mach. Intell. 44 (9), 5700–5714.

Yu, J., Hu, Y., Ning, X., Qiu, J., Guo, K., Wang, Y., Yang, H., 2017. Instruction driven cross-layer CNN accelerator with winograd transformation on FPGA. In: 2017 International Conference on Field Programmable Technology (ICFPT), IEEE, pp. 227–230.

Zagoruyko, S., Komodakis, N., 2017. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: International Conference on Learning Representations (ICLR).

Zeng, H., Xie, X., Cui, H., Zhao, Y., Ning, J., 2020. Hyperspectral image restoration via cnn denoiser prior regularized low-rank tensor recovery. Comput. Vis. Image Underst. 197, 103004.

Zeng, R., Lu, Z., Wang, J., Song, J., 2022. Error Correction Coding for One-Bit Quantization With CNN-Based AutoEncoder. IEEE Commun. Lett. 26 (8), 1814–1818.

Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H., 2018. Deep mutual learning, in. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4320–4328.

Zhao, R., Song, W., Zhang, W., Xing, T., Lin, J.-H., Srivastava, M., Gupta, R., Zhang, Z., 2017. Accelerating binarized convolutional neural networks with software-programmable FPGAs. In: Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, pp. 15–24.

Zhou, A., Yao, A., Guo, Y., Xu, L., Chen, Y., 2017. Incremental network quantization: Towards lossless CNNs with low-precision weights. In: International Conference on Learning Representations (ICLR).

Zhu, M., Gupta, S., 2017. To prune, or not to prune: exploring the efficacy of pruning for model compression, arXiv preprint arXiv:1710.01878.