Noise transfer for unsupervised domain adaptation of retinal OCT images

Moritz Haderer
Advisors - Taha Emre, Hrvoje Bogunovic
Practical work in AI
Johannes Kepler Universität Linz

Abstract

In optical coherence tomography (OCT), high-resolution scans of patients' eyes enable professionally trained ophthalmologists to diagnose a variety of eye diseases. Machine learning techniques have shown significant potential across many image analysis tasks, leveraging information that might be overlooked by physicians and integrating it into their diagnostic routines. Semantic segmentation models applied to OCT imaging helps detect fluid accumulations in an efficient manner and supports medical experts in delivering care. Training machine learning models in OCT image analysis is difficult, as labelled datasets are scarce and there exist differences among OCT device manufacturers that make generalization challenging. Singular Value Decomposition-based Noise Adaptation (SVDNA) is a method has demonstrated promising results in improving generalization across devices and it does so requiring labels only for one device's scans. Using the RETOUCH dataset, SVDNA is tested in its generalisation capabilities and is compared to the fully supervised setting, which requires more labelled data. Ablation studies are carried out to further dissect the method.

- **What is the central question?**
  How can the shift between domains caused by differences in OCT imaging hardware be bridged in machine learning models while labeled data is limited?
- **Why is this question important?**
  If the goal is to create a model that can solve the task of fluid accumulation segmentation in OCT images with human-like competence, then domain adaptation is a crucial part of the puzzle. A model that is able to extract general features across domains is the simpler and more elegant solution compared to training multiple models. Additionally, labeled data is scarce, hence domain adaptation to unlabeled domains is an important research direction.
- **What evidence/data (variables) are needed to answer this question?**
  A performance comparison of models without the method in question (control group) and models where the method in question has been applied (intervention group). Additionally, a baseline performance of a naive model should be available. In the realm of unsupervised and semi-supervised learning, a fully supervised model can serve as a "ceiling", which should, ideally, be attained.
- **What methods are used to get this evidence/data?**
  Training several models on the same training set in a deterministic manner. Then, testing on an unseen test set can yield the information necessary to compare performances. More specifically, we want to investigate whether a model trained on two labeled source domains can perform equal to a model trained on three labeled source domains if the method in question (SVDNA) is applied to the former. Applying the method in question successfully would let us omit an entire labeled domain dataset without losing performance.

- **What analyses must be applied for the data to answer the central question?**
  Before anything is analyzed, a suitable metric must be chosen that is able to reflect the performance we want to measure. This metric is then used to quantify the differences in model performances, most often done by comparing the means of said metric. As long as the test set is large enough, comparing the mean performance of a control and an intervention group can yield insight into the method's benefits. Furthermore, ablation studies can give further insight into effect of a certain method by investigating which parts are beneficial.
- **What evidence/data (values for the variables) were obtained?**
  The Dice score of segmentation predictions vs. ground truth images was calculated across several models and several prediction tasks. As there are three different domains available, six models were trained: Every pair of two domains as labeled source domains and applying SVDNA to gain knowledge of an unlabeled target domain (e. g. Domain 1, Domain 2 supervised + Domain 3 unsupervised). Then, every pair of two domains but without using the method in question. Furthermore, a fully supervised model using all three domains and their labels.
- **What were the results of the analyses?**
  The performance results of the method showed an increase in performance whenever SVDNA was used. The fully supervised ceiling model was not quite topped, but performance came close in some sub tasks. Unfortunately, however, The structure of the dataset posed a problem that somewhat clouds the results. A potential difference in class distribution between the training and the test set lowered the average performance of every model by a significant amount, since all of the models were trained on data where fluid accumulation was, in fact, present. In the test set, more than 50% of all data can be assumed to have empty labels (where no fluid accumulation is present).

  The ablation study performed showed that a less central part of the method, the histogram matching performed at the end of SVDNA, affects performance to a significant degree. Noise adaptation, which the method is named after, even appeared detrimental in one case and could only yield beneficial results if used in conjunction with histogram matching.
- **How did the analyses answer the central question?**
  The analyses showed that, in the realm of OCT imaging, differences in domains are, to some degree, constituted by the character and tone present in a given domain. By extracting these particularities of one domain and adding them to another, a model can adapt to the differences in pixel distribution caused by them. The analyses hence showed that SVDNA is a method capable of somewhat bridging the gap between domains in a way that does not require more labeled data.
- **What does this answer tell us about the broader field?**
  A vision model's performance is directly linked to the distribution of values it is able to associate with a given outcome. When a shift in distribution takes place, performance quickly deteriorates. The answer tells us that creative image augmentation is a powerful tool to confront a model with a broader range of scenarios to learn proper behavior in.

1 Introduction

Optical coherence tomography is an imaging technique that is commonly used in ophthalmology. It captures high resolution 3D scans of the eye without being invasive for the patient and has become the state of the art in diagnostic imaging for retinal diseases. Many retinal diseases occur in conjunction with fluid accumulation in various layers of the retina. Measuring these fluid volumes can be a challenging task for ophthalmologists. OCT scans are noisy and prone to distortions due to eye movements of the patient. Some fluid biomarkers are very hard to discern even by experts. Additionally, each OCT scan yields a volume, rather than a 2D image, making the task of fluid detection even more challenging. Machine learning methods have shown immense success in medical image analysis tasks and continue to evolve in sophistication. Semantic segmentation is a dense prediction problem in which a machine learning model learns to fill regions of interest in an input image. The application of computer vision methods to OCT scans shows encouraging results in fluid accumulation segmentation and is a promising avenue for improving patient care. Despite all the success, ophthalmological image analysis, especially concerning OCT images, bears unsolved problems. With annotation being a difficult task requiring expert knowledge, labelled datasets are

scarce. Furthermore, OCT imaging device manufacturers face a trade-off in resolution, scan time, and signal-to-noise ratio in building the scanners. As a consequence, different device brands' scans exhibit variation in image quality and noise characteristics. These differences pose a problem that, in the machine learning world, is known as a "domain shift". Training a machine learning model on one device brand's domain does not guarantee generalization to other devices' images. In this report, SVDNA Koch et al. [2022] is described and tested, a novel technique which tackles two problems: 1) the scarcity of annotated datasets in OCT imaging and 2) extending model generalization capabilities to several domains due to differences in OCT devices. By using singular value decomposition and histogram matching, the characteristics of unlabelled target images are transferred to labelled source images, enabling a model to learn multiple domains' particularities from a single source domain in a supervised manner. Extensive experimentation is carried out on the RETOUCH dataset Bogunovic et al. [2019] to test the viability of this method on a benchmark dataset. The main hypothesis is that a model trained on two of three OCT scan domains using SVDNA should perform on par with a fully supervised model. This hypothesis was tested on all domain pairings and an ablation study was carried out on the best performing duo. While SVDNA did increase the models' performance across all experiments, improvements also depended on the respective target domain's characteristics, the noisiest domain benefiting the most.

2 Dataset and Augmentations

In ophthalmology, retinal images from OCT devices can be used to diagnose, track, and, hopefully, prevent several eye diseases such as glaucoma, age-related macular degeneration (AMD), and diabetic retinopathy. There are several fluid biomarkers that give insight into disease presence and progression. On OCT images, three types of fluid are clinically distinguishable and considered relevant for clinical diagnosis: Intraretinal fluid (IRF), which has been found to be a strong predictor for vision loss, Subretinal fluid (SRF), which gives an indication of a possibly favorable disease progression in AMD, and Pigment epithelial detachment (PED), which is considered a major indicator for progressive disease activity.

Despite the large amount of OCT images coming out of ophthalmological practise, the availability of professionally labeled datasets remains limited, and there is a lack of standardized methods for comparing image analysis tools specialized in this type of imaging. The RETOUCH dataset was developed to address this issue.

2.1 RETOUCH

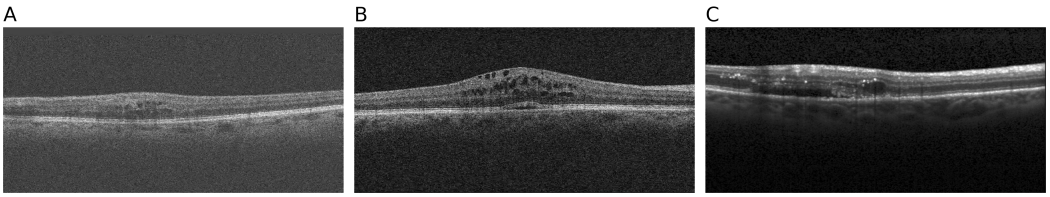
RETOUCH is a large, professionally annotated set of OCT volumes acquired using three different device brands. Originally published in the context of a MICCAI grand challenge in 2017, it was collected to boost progress in machine learning-based approaches in ophthalmological image analysis research. It is the first benchmark dataset for classification and segmentation of fluid biomarkers in retinal images. The dataset consists of 112 samples, each of which is a 3D volume of an OCT scan of a single patient. The volumes have been collected by the Medical University in Vienna (MUW), the Erasmus University Medical Center (ERASMUS), and the Radbound University Medical Center (RUNMC). Each volume was labelled by opthalmologists at MUW or RUNMC and was randomly split into training and test set in a ratio of approximately 60%:40%.

While the output of OCT devices consists of 3D volumes, the method investigated in this report is defined for 2D slices of said volumes. Fortunately, the team at the OPTIMA group provided a pre-processed version of the dataset, where volumes had already been sliced into 2D images, flattened, and denoised - all of which are common pre-processing steps in OCT imaging. Depending on the device brand used for the scan, a single volume can be divided into a varying number of 2D B-scans. Once the volumes are sliced into 2D scans the total amount of samples in the training set and the test set is 6936 and 4270, respectively.

Note that among the 6936 images in the training set, only 3385 have corresponding label masks, because scans without fluid accumulation present do not require labels. In the test portion of the dataset, however, all images have labels. As a consequence of this, a model trained on the training set without additionally added empty labels would learn a different mapping than the test case requires, since the ratio of background to fluid mask increases in the test set. This will be addressed in more detail in section 3.1.3 and 5.

The RETOUCH dataset contains OCT scans from three different brands of OCT devices: Spectralis, Topcon, and Cirrus. Each device produces images with underlying diffferences: Varying noise levels, image quality, and sharpness lead to a difference in pixel distributions that is large enough to have machine learning algorithms struggle to generalize from one device's scans to another. In Figure 1 the differences between brands are shown. When creating a scan using a Spectralis device, several B-scans are averaged at certain positions to increase image resolution. When slicing volumes into B-scans, Spectralis devices yield 49 single images, while Topcon and Cirrus each yield 128.

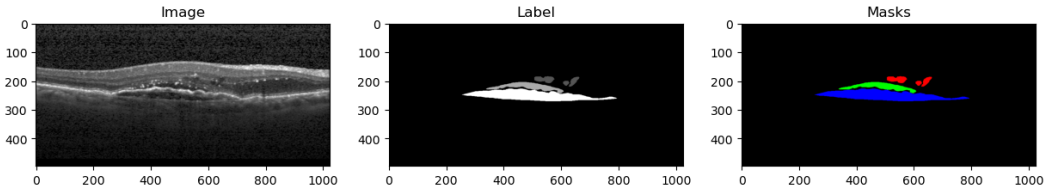
Figure 1: Three scanner brands compared. A) Topcon B) Cirrus C) Spectralis. Note the difference in noise levels and sharpness of edges.



2.2 Data pipeline and augmentations

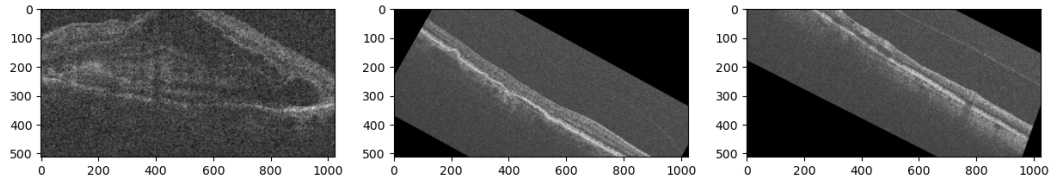
To effectively use the image data for training, a robust data pipeline and augmentation process was established using MONAI Cardoso et al. [2022] for augmentations and PyTorch Paszke et al. [2019] for dataloading. Since the data was provided in a form that did not require volume slicing, retinal layer segmentation, layer segmentation, and denoising, a large chunk of the usual OCT pre-processing pipeline could be omitted. The pre-processed B-scans have a resolution of 1024x496. The labels are provided as gray-scale segmentation masks with the different biomarkers being represented by pixel values 50, 100 or 150, while the background is always zero. As a first step in the pipeline, the labels were separated into four channels, three for each fluid biomarker and one for the background. Figure 2 shows an OCT scan and its original labels as well as the labels converted into multi-channel masks.

Figure 2: Original image, provided labels and channel-separated segmentation masks. In the channel-separation, the model will predict zeros (background) and ones separately for each channel representing a biomarker.



A number of image augmentation steps was included to boost the robustness of the trained models. Using MONAI, augmentation was carried out on the image and the masks together. Transformations like rotation or axis flipping should only be applied during model training, as this boosts robustness. There were some transforms, however, that were necessary for both training and evaluation: the pixel values were normalized to the range [0, 1] and the images were resized to 1024 x 512 to comply with model requirements (Section 3.1.2). Then, a series of image augmentations was applied only to the training data: a randomized zoom, randomized axis flipping, and randomized affine transforms like translation, shearing, and rotation. The probability of each randomized transform in the pipeline happening was set to 0.3. Their respective parameters were determined visually and were kept in a "reasonable" range. Specifics can be found in the appendix. Figure 3 shows the augmentations applied to training data.

Figure 3: Three randomly sampled images with training augmentations applied. The leftmost image is zoomed-in, the other two are translated and rotated.



3 Methods

SVDNA The method was introduced by Koch et al. [2022] as a way to mitigate domain shifts between OCT scanner brands. The authors hypothesize that the difference in noise signatures is a decisive factor in the domain differences and devised an algorithm to mitigate these differences. SVDNA is a model agnostic approach and an extension to any existing image augmentation pipeline. By using Singular Value Decomposition (SVD) to decompose images into their singular components, the authors state that it is possible to extract the noise of an image by taking the k last singular values. Extracted noise levels from unlabeled target domains transfer noise characteristics to labelled source domain images. Finally, a histogram matching step is performed, bringing the source image's distribution closer to the target image's. In other words, SVDNA helps align the distributions of source domain and target domain data. This has been shown to effectively extends a model's capabilities to generalize to domains that were included as target domains and improves model performance across diverse datasets.

SVDNA was included in the data processing pipeline before any other augmentation steps. Upon construction of a dataset object, a source domain and target domain(s) must be chosen. Only source domain images are required to have annotations. As a data sample from the source domain is selected, another domain to sample from is chosen with a probability of $\frac{1}{p}$, where p is the number of domains in the dataset. If the source domain is chosen, no SVDNA is performed. If a target domain is chosen, a value k for the magnitude of noise to be transferred is sampled from the range [30, 50]. This range was empirically established by the authors. After noise adaptation, histogram matching was performed. Algorithm 1 outlines the steps involved in the SVDNA approach again, emphasizing noise transfer and histogram matching to achieve domain adaptation.

Algorithm 1 SVDNA [Koch et al., 2022]

Let $U\Sigma V^T$ be the singular value decomposition of the respective images im_{source}, $\text{im}_{target} \in \mathbb{R}^{n\times n}$ and k the noise transfer threshold.

$\text{im}_{source} = U_s\Sigma_s V_s^T, \quad \text{im}_{target} = U_t\Sigma_t V_t^T,$
$U_r \leftarrow u_s^1, ..., u_s^k, u_t^{k+1}, ..., u_t^n$
$\Sigma_r \leftarrow \text{diag}(\sigma_s^1, ..., \sigma_s^k, \sigma_t^{k+1}, ..., \sigma_t^n)$
$V_r^T \leftarrow {v_s^1}^T, ..., {v_s^k}^T, {v_t^{k+1}}^T, ..., {v_t^n}^T$
$\text{Im}_{noised} \leftarrow U_r\Sigma_r V_r^T$
$\text{Im}_{clipped} \leftarrow$clip_values_to_interval($\text{Im}_{noised}$,$[0, 255]$)
$\text{Im}_{restyled_final} \leftarrow$histogram_matching(source=$\text{Im}_{clipped}$, target=im_{target})

3.1 Training and experiments

3.1.1 Frameworks

PyTorch An open-source deep learning framework developed by Facebook's AI Research lab. Known for its dynamic computational graph, automatic differentiation capabilities, and intuitive inter-

face, PyTorch allows developers to build, train, and deploy neural networks. PyTorch's flexibility and strong community support make it popular among researchers for prototyping and experimentation. Its extensive ecosystem, including libraries for vision, natural language processing, and reinforcement learning, has established PyTorch as a leading framework in both academic research and industry applications. [Paszke et al., 2019]

PyTorch Lightning A high-level interface for PyTorch that simplifies the training process for deep learning models. It provides a lightweight skeleton for organizing code, automating training tasks, and enabling reproducibility. By abstracting away boilerplate code, experiments become more organized and transparent. Lightning handles complex aspects like distributed training, gradient accumulation, and mixed precision, which makes it useful for scaling research projects without noticeable effort. In this project, it was used to organize training, validation, testing, logging, and model checkpointing. [Falcon and The PyTorch Lightning team, 2019]

MONAI Medical Open Network for AI was introduced to streamline deep learning workflows in healthcare applications. Built on top of PyTorch, MONAI provides a comprehensive set of tools and utilities tailored for medical imaging tasks. It was used in establishing an efficient image augmentation pipeline and offers a variety of loss functions used in image segmentation. [Cardoso et al., 2022]

Weights and Biases Weights and Biases (WandB) is a tool widely used for logging and tracking experimental results. By linking logged data to an online server, results can be visualized dynamically. Apart from logging losses and metrics, entire tables, images and model checkpoints can be recorded. The service also provides options for hyper-parameter search and integrates nicely with Pytorch Lightning. In this report, WandB was used to track model training and to record metrics like Dice scores, precision, and recall. [Biewald, 2020]

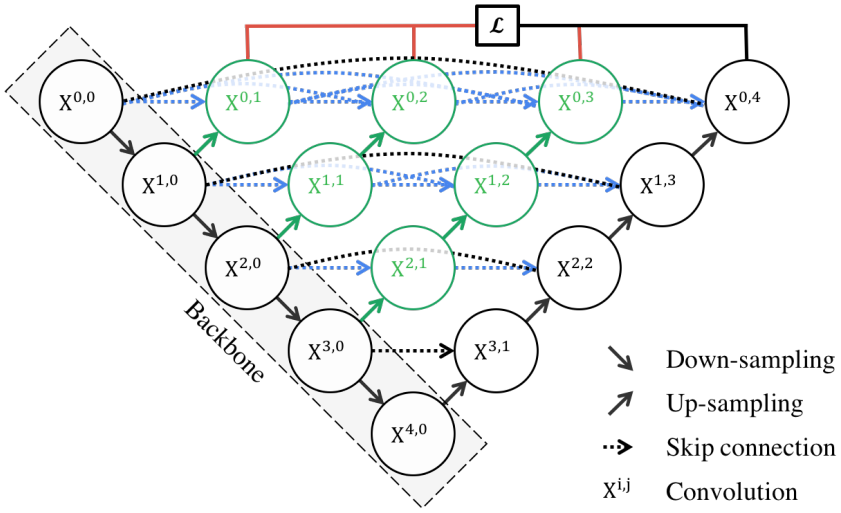
3.1.2 Network

A UNet++ Zhou et al. [2018] was used for the segmentation task. The model implementation was imported from "segmentation-models-pytorch" [Iakubovskii, 2019], which provides a wide variety of pretrained segmentation model weights. The UNet++ architecture extends the traditional UNet by adding dense skip connections and nested, densely connected convolutional layers. These changes aim to improve feature fusion and representation learning. Unlike the regular UNet, UNet++ boasts a more complex structure with multiple intermediate layers. These layers can be said to refine the extracted features before a final upsampling, which leads to more precise segmentation results. The UNet++ was chosen due to its superior performance in dense segmentation tasks and increased reconstruction resolution. The model used for this task is composed of an ImageNet-pretrained ResNet18 encoder and a decoder with feature map channels [1024, 512, 256, 128, 64]. The implementation also includes the Channel-Space attention mechanism (SCSE): an addition to convolutional neural networks that increases global feature extraction by paying special attention to channel-wise as well as kernel-wise features. ImageNet pretraining aids in biomedical imaging by leveraging learned features relevant to textures, edges, and shapes, enabling better generalization. In total, the model is relatively large with 41 million tunable parameters.

3.1.3 Training

A grand total of 12 models were trained using the hyper-parameters as described in Koch et al. [2022]. Training was carried out using an A6000 GPU with 50GB VRAM on the computing cluster of the OPTIMA Lab at the Medical University of Vienna. Although almost all training specifics could be adopted from the original paper, some changes had to be made due to the difference in task details between the originally used dataset and the RETOUCH dataset. Specifically, the difference regards how testing was carried out. In the original paper, testing was largely done using five-fold cross validation on the training set. As no "healthy" scans - scans with empty segmentation masks - were included in the training set, the model's ability to discern when to only predict background was impaired. This also means that the ratio of background to biomarkers pixels was the same for training and testing in the SVDNA paper implementation. In the implementation made for this report, testing is done on RETOUCH's official test portion and there is no guarantee that "healthy" scans will be excluded. If there is, indeed, a significant percentage of "healthy" scans present in the test set, a model needs to be able to discern "healthy" from "unhealthy" during testing. Upon request, the

Figure 4: Architecture of the UNet++. Dense skip connections are shown by dashed arrows. In this figure, there are four upsampled output layers. By carrying out multiple loss calculations, deep supervision Li et al. [2022] can be incorporated into UNet++ training. The figure was taken from the original paper Zhou et al. [2018].



SVDNA paper's authors also specified that they excluded the background from loss calculations during training. In this report's implementation, it had to be included in order to force the model to better learn how background should be distributed. The loss function used in the paper is the a combination of binary cross entropy and Dice loss, while here, BCE + generalized Dice loss was adopted in due to its class imbalance-correcting capabilities. The formula for the generalized Dice loss is shown in Equation 1.

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{N} w_i p_i g_i}{\sum_{i=1}^{N} w_i (p_i + g_i)} \tag{1}$$

N is the number of classes, p_i is the predicted probability for class i, g_i is the ground truth binary label for class i, and w_i is the class weighting factor, which can be calculated as $w_i = (\sum_{j=1}^{N} g_j)^{-2}$.

Table 1: Parameters used for training of all models.

Parameter	**Value**
Learning Rate	1×10^{-4}
LR Scheduler	Reduce on Plateau Decay factor: 0.003
Optimizer	AdamW Weight decay: 0.003
Batch Size	8
Epochs	130
Loss Function 1	Generalized Dice Loss
Loss Function 2	Binary Cross Entropy BCE weighting: 0.5

3.1.4 Experiments

As a first experiment, empty labels were generated for each of the "healthy" samples and a fully supervised UNet++ was trained using a combination of dice loss and binary cross entropy. The hypothesis was that the model's generalization capabilities may benefit from learning the underlying distribution of healthy and unhealthy samples and, thus, perform better in testing. However, the

experiment showed that this combination of loss functions was unable to learn how to solve this segmentation task due to the severe class imbalance. Dice loss fluctuated between 0 and 1 without convergence and after 50 epochs training was halted. Possible improvements could have been made by using a combination generalized dice loss and focal loss, as they both include a form of class imbalance correction by weighting the respective loss by the proportion of pixels in a given class.

In further experiments, the combination of generalized dice loss and binary cross entropy was adopted. Also, the previously generated empty labels were excluded from the training set. This effectively reduced the training set size by more than 50 %, but corrected the severe imbalance between background and biomarker pixels enough for training to converge.

The primary hypothesis to test was that SVDNA should effectively enable a model to bridge the distribution shift between labelled source and unlabelled target domains. Bridging the domain shift means that a model trained in a supervised manner on two of three domains, but with SVDNA enabled, would perform equally to a model trained in a fully supervised setting in all three domains. Training on two of three domains with SVDNA means that no annotations are required for the target domain, as merely transferring noise characteristics and histogram matching of said domain to the source should suffice to extend the model's generalization capabilities to the target domain.

As a baseline, a model with only ImageNet-pretrained encoder weights and no decoder training is included. The fully supervised model including all three domains (Spectralis, Topcon, Cirrus) represents the ceiling, as none of the other experiments should surpass it.

4 Results

As a metric for comparing experimental results, the Dice (F1) score was used. Table 2 presents the main results, comparing the scores for each experiment and each biomarker class. The mean is calculated over all fluid types. Testing has been carried out on the official test set of the RETOUCH dataset. Upon running the first tests, it turned out that there is a severe class distribution shift from the training set to the test set. Since the test set results show a highly abnormal distribution (See: Figure 5), the decision to also include testing results on the validation set containing all three domains (abbreviated by "V" in Table 2) was made.

In all experiments, an increase in Dice score was observed whenever SVDNA was included in the image augmentation pipeline. The most notable improvement from control to intervention is in the case "Spectralis + Cirrus". While models with SVDNA performed better overall, the fully supervised ceiling could not be reached in any of the experiments. The "Spectralis + Cirrus" model mostly closed the gap to the supervised model with a difference in Dice score of only 0.006.

While the mean results on the test set point to a beneficial effect caused by using SVDNA, the standard deviations are very large and clearly raise doubt as to how reliable the means are. The top panel in figure 5 shows a very uneven distribution of Dice scores on the test set. Between 69% (Supervised model) and 73% (S+C without SVDNA) of all Dice scores are zero or negligibly close to zero. The rest is spread out in the upper third, representing relatively good segmentation predictions. Metrics like precision ($\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$) and recall ($\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$) are reported in the appendix (Table 7.2), as they provide more insight into distribution of false positives and false negatives. Precision on the official test set is between 50 % and 100 % higher than recall in all cases.

Since SVDNA brought the largest increase in performance when the domain "Topcon" was the target, this case has been used for an ablation study. We see that the *noise adaptation* step of the SVDNA method yielded large improvement only in the case of SRF. In PED, the change was relatively small and in IRF it even worsened due to noise adaptation. Histogram matching, however, outperformed even the full SVDNA method by a small margin in two of three tasks, with mean performance of only histogram matching being almost equal to the full method (0.211 vs. 0.210).

5 Discussion

The validation results indicate that the training process was successful, and the models' performance aligns with expectations. While the top panel in Figure 5 reveals a heavily biased distribution of scores, this bias is relatively homogeneous across all models, suggesting that the means can still provide valuable insights into the average model performance. The Y-axis reaches a maximum of

Table 2: Results of the experiments on SVDNA performance. Mean Dice (F1) scores are reported with standard deviations in brackets. Reported results are for the official RETOUCH test set (T) as well as for the validation set containing all three domains (V). Task abbreviations: IRF = Intraretinal fluid, SRF = Subretinal fluid, PED = Pigement epithelial detachment. In column "Model", abbreviations are S : Spectralis, C : Cirrus, T : Topcon. In all experiments, SVDNA boosted the Dice score of the respective task.

Model		**IRF**	**SRF**	**PED**	**mean**
ImageNet pretraied	T	0.002 (0.01)	0.005 (0.02)	0.003 (0.01)	0.003 (0.01)
	V	0.01 (0.01)	0.0 (0.0)	0.017 (0.02)	0.009 (0.01)
S+T (No SVDNA)	T	0.247 (0.35)	0.173 (0.32)	0.143 (0.27)	0.188 (0.31)
	V	0.736 (0.15)	0.738 (0.22)	0.667 (0.27)	0.713 (0.21)
S+C (No SVDNA)	T	0.237 (0.34)	0.103 (0.24)	0.132 (0.27)	0.157 (0.28)
	V	0.723 (0.2)	0.627 (0.42)	0.732 (0.25)	0.694 (0.29)
C+T (No SVDNA)	T	0.245 (0.34)	0.174 (0.33)	0.155 (0.29)	0.191 (0.32)
	V	0.748 (0.18)	0.852 (0.13)	0.798 (0.19)	0.8 (0.17)
S+T (SVDNA)	T	0.252 (0.35)	0.182 (0.33)	0.154 (0.29)	0.196 (0.32)
	V	0.743 (0.14)	0.821 (0.16)	0.767 (0.17)	0.777 (0.16)
S+C (SVDNA)	T	0.263 (0.36)	0.194 (0.35)	0.172 (0.3)	0.21 (0.34)
	V	0.778 (0.11)	0.842 (0.14)	0.803 (0.16)	0.807 (0.14)
C+T (SVDNA)	T	0.254 (0.35)	0.180 (0.34)	0.164 (0.3)	0.199 (0.33)
	V	0.766 (0.13)	0.865 (0.11)	0.816 (0.16)	0.815 (0.13)
Fully supervised	T	0.265 (0.36)	0.197 (0.35)	0.185 (0.32)	0.216 (0.34)
	V	0.828 (0.05)	0.898 (0.11)	0.875 (0.11)	0.867 (0.09)

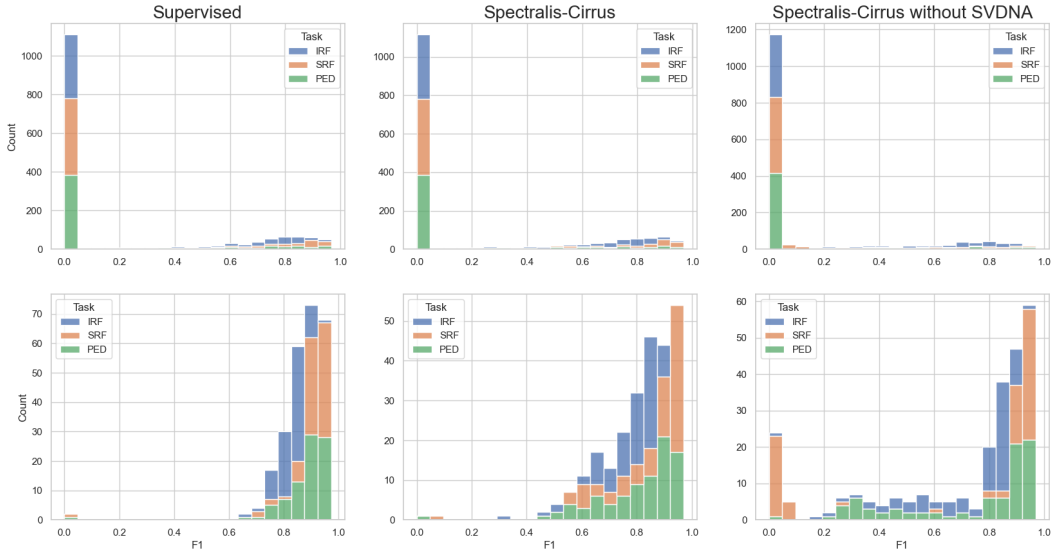
Table 3: **Ablation:** Mean (SD) of dice (F1) score for ablations of best performing data subset. NA = Noise adaptation, HM = Histogram matching.

Model	**IRF**	**SRF**	**PED**	**mean**
Spectralis + Cirrus (no SVDNA)	0.237 (0.34)	0.103 (0.24)	0.132 (0.27)	0.157 (0.28)
Spectralis + Cirrus (NA)	0.222 (0.34)	0.187 (0.34)	0.138 (0.27)	0.182 (0.31)
Spectralis + Cirrus (HM)	0.258 (0.35)	**0.198 (0.35)**	**0.175 (0.31)**	**0.211 (0.34)**
Spectralis + Cirrus (SVDNA)	**0.262 (0.36)**	0.194 (0.35)	0.172 (0.3)	0.210 (0.34)
Supervised training	0.265 (0.36)	0.197 (0.35)	0.185 (0.32)	0.216 (0.34)

1601, derived from the number of samples in the test set (4270) divided by the batch size (8) and multiplied by the number of fluid biomarkers (3). The zero-bias shown in the figure is reflected in the results, where the means are pulled towards zero and the standard deviations are inflated.

The large proportion of Dice scores of zero, approximately 70 %, may stem from the class distribution in the official RETOUCH test set. In the training set, 51 % of all images lack signs of fluid accumulation and therefore carry no labels. This may also hold true for the test set. According to the RETOUCH paper Bogunovic et al. [2019], a 60-40 random split was performed. Under this assumption, the prevalence of background class labels increases significantly, leading any model to predict excessive false positives. However, this hypothesis is contradicted by the precision and recall results. With precision generally higher than recall, this indicates a higher false negative rate, where

Figure 5: **Top:** Histogram of Dice score distributions of three models on the independent test set. Dice score of zero indicates no or negligible overlap between prediction and ground truth. There were 1111, 1116, and 1173 samples with a Dice score close to zero in the "Supervised" model, in the "Spectralis-Cirrus" model, and in the "Spectralis-Cirrus without SVDNA" model, respectively. This amounts to almost three quarters of samples in all three classes.
Bottom: Histogram of Dice score distributions of the same models on the validation set containing all three domains.



the model predicts too much background. This scenario would only occur if the test set contains a higher proportion of fluid accumulation per scan and cannot be assumed to be true.

Another hypothesis is that the models primarily learn to recognize the presence of background pixels, which leads to them frequently predicting background instead of fluid biomarkers. They consequently underestimate the extent of segmentation required. At the same time, the test set contains an immense number of black (background) labels. The models have not encountered purely black labels during training, so they tend to predict some degree of segmentation in these areas. Consequently, this results in many false negatives, as the models miss out in dense segmentation zones, and a significant number of false positives due to the excessive amount of black images in the test set. This dual tendency to both underpredict in areas requiring segmentation and overpredict in regions without significant features highlights the need for further model refinement and dataset balancing.

In real-world applications, a fluid segmentation model is likely used on samples suspected of disease markers. Therefore, a model specialized in segmenting existing fluid accumulations, rather than ignoring a lack thereof, would not be detrimental. Despite the low reported Dice scores, random predictions from the model, as illustrated in the appendix, show accurate localization of fluid accumulations, suggesting practical utility in relevant scenarios.

Additionally, the ablation study highlights the specific contributions of various components within the SVDNA pipeline. Noise adaptation significantly improved performance for SRF, had negligible effects on PED, and worsened performance for IRF. Conversely, histogram matching consistently benefited all fluid types and sometimes outperformed the full SVDNA method. This implies that while the overall SVDNA approach is effective, certain elements like histogram matching might be more critical and should be prioritized in future implementations to optimize model performance.

Notable limitations include the denoising capabilities of the SVDNA method. Training a model on a high-noise domain like Topcon, noise transfer from a low-noise domain like Spectralis may not have a large effect. As the experiments show, noise transfer from Topcon enhances the models' capabilities, but the opposite does not hold. Further research could address the possibility of using SVDNA-like methods on learned representation, instead of raw images. If domain-specific features are relevant enough in representations to impede domain-invariant applications, then we might also be able to

extract and adapt these features. The current implementation of SVDNA is not GPU-enabled and a significant gain in speed could be made by creating an implementation using PyTorch.

6 Conclusion

In this report, the method SVDNA was investigated in its capabilities for enhancing generalization in machine learning models across OCT image domains. By training models on two of three domains and comparing the results to a fully supervised model, a comparative ceiling, its performance could be gauged. While the results indicate an overall improvement in dice scores over all models using SVDNA, possible class distribution differences between the training and the test set have introduced a potential bias. SVDNA has been shown to be a relatively light-weight, modular addition to any image augmentation pipeline and it is a flexible way to exploit unlabeled data of another domain. Ablation studies showed a surprisingly low impact of the computationally expensive noise adaptation part and demonstrated the importance of the histogram matching for the method.

References

Lukas Biewald. Experiment tracking with weights and biases, 2020. URL `https://www.wandb.com/`. Software available from wandb.com.

Hrvoje Bogunovic, Freerk Venhuizen, Sophie Klimscha, Stefanos Apostolopoulos, Alireza Bab-Hadiashar, Ulas Bagci, Mirza Faisal Beg, Loza Bekalo, Qiang Chen, Carlos Ciller, Karthik Gopinath, Amirali K. Gostar, Kiwan Jeon, Zexuan Ji, Sung Ho Kang, Dara D. Koozekanani, Donghuan Lu, Dustin Morley, Keshab K. Parhi, Hyoung Suk Park, Abdolreza Rashno, Marinko Sarunic, Saad Shaikh, Jayanthi Sivaswamy, Ruwan Tennakoon, Shivin Yadav, Sandro De Zanet, Sebastian M. Waldstein, Bianca S. Gerendas, Caroline Klaver, Clara I. Sanchez, and Ursula Schmidt-Erfurth. RETOUCH: The Retinal OCT Fluid Detection and Segmentation Benchmark and Challenge. *IEEE Transactions on Medical Imaging*, 38(8):1858–1874, August 2019. ISSN 0278-0062, 1558-254X. doi: 10.1109/TMI.2019.2901398. URL `https://ieeexplore.ieee.org/document/8653407/`.

M. Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, Vishwesh Nath, Yufan He, Ziyue Xu, Ali Hatamizadeh, Andriy Myronenko, Wentao Zhu, Yun Liu, Mingxin Zheng, Yucheng Tang, Isaac Yang, Michael Zephyr, Behrooz Hashemian, Sachidanand Alle, Mohammad Zalbagi Darestani, Charlie Budd, Marc Modat, Tom Vercauteren, Guotai Wang, Yiwen Li, Yipeng Hu, Yunguan Fu, Benjamin Gorman, Hans Johnson, Brad Genereaux, Barbaros S. Erdal, Vikash Gupta, Andres Diaz-Pinto, Andre Dourson, Lena Maier-Hein, Paul F. Jaeger, Michael Baumgartner, Jayashree Kalpathy-Cramer, Mona Flores, Justin Kirby, Lee A. D. Cooper, Holger R. Roth, Daguang Xu, David Bericat, Ralf Floca, S. Kevin Zhou, Haris Shuaib, Keyvan Farahani, Klaus H. Maier-Hein, Stephen Aylward, Prerna Dogra, Sebastien Ourselin, and Andrew Feng. MONAI: An open-source framework for deep learning in healthcare, November 2022. URL `http://arxiv.org/abs/2211.02701`. arXiv:2211.02701 [cs].

William Falcon and The PyTorch Lightning team. PyTorch Lightning, March 2019. URL `https://github.com/Lightning-AI/lightning`.

Pavel Iakubovskii. Segmentation models pytorch. `https://github.com/qubvel/segmentation_models.pytorch`, 2019.

Valentin Koch, Olle Holmberg, Hannah Spitzer, Johannes Schiefelbein, Ben Asani, Michael Hafner, and Fabian J. Theis. Noise transfer for unsupervised domain adaptation of retinal OCT images. volume 13432, pages 699–708. 2022. doi: 10.1007/978-3-031-16434-7_67. URL `http://arxiv.org/abs/2209.08097`. arXiv:2209.08097 [cs, eess].

Renjie Li, Xinyi Wang, Guan Huang, Wenli Yang, Kaining Zhang, Xiaotong Gu, Son N. Tran, Saurabh Garg, Jane Alty, and Quan Bai. A Comprehensive Review on Deep Supervision: Theories and Applications, July 2022. URL `http://arxiv.org/abs/2207.02376`. arXiv:2207.02376 [cs].

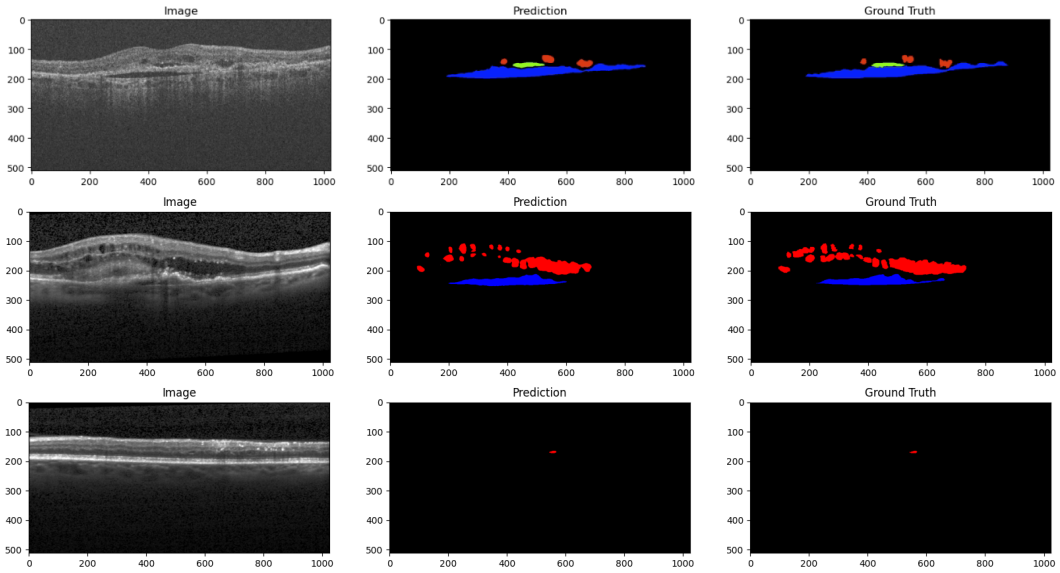
Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library, December 2019. URL `http://arxiv.org/abs/1912.01703`. arXiv:1912.01703 [cs, stat].

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. UNet++: A Nested U-Net Architecture for Medical Image Segmentation, July 2018. URL `http://arxiv.org/abs/1807.10165`. arXiv:1807.10165 [cs, eess, stat].

7 Appendix

7.1 Predictions

Figure 6: Model Supervised; Training set S+C+T



7.2 Further results

Table 4: Augmentation parameters

Augmentation	**Parameters**
Zoom	Magnification 1x - 1.5x
Rotation	(-15, 15) degrees
Shearing	X-Axis: (-0.2, 0.2)
	Y-Axis: (0, 0)
Translation	X-Axis: (-100, 100)
	Y-Axis: (0, 0)

Table 5: Full results

Task	Model	Accuracy	Precision	Recall
IRF	Untrained	0.431	0.001	0.078
	Cirrus	0.998	0.288	0.236
	Topcon	0.998	0.271	0.232
	Spectralis	0.997	0.204	0.210
	Topcon-Cirrus-noSVDNA	0.998	0.295	0.230
	Spectralis-Topcon-noSVDNA	0.998	0.250	0.267
	Spectralis-Cirrus-noSVDNA	0.998	0.277	0.229
	Spectralis-Topcon	0.998	0.284	0.251
	Topcon-Cirrus	0.998	0.296	0.242
	Spectralis-Cirrus	0.999	0.285	0.268
	Supervised	0.999	0.290	0.265
PED	Untrained	0.640	0.001	0.027
	Cirrus	0.997	0.229	0.127
	Topcon	0.996	0.223	0.112
	Spectralis	0.994	0.146	0.086
	Topcon-Cirrus-noSVDNA	0.997	0.225	0.136
	Spectralis-Topcon-noSVDNA	0.996	0.232	0.126
	Spectralis-Cirrus-noSVDNA	0.996	0.205	0.115
	Spectralis-Topcon	0.996	0.227	0.136
	Topcon-Cirrus	0.997	0.223	0.148
	Spectralis-Cirrus	0.997	0.231	0.154
	Supervised	0.997	0.236	0.170
SRF	Untrained	0.773	0.003	0.050
	Cirrus	0.999	0.225	0.169
	Topcon	0.998	0.220	0.140
	Spectralis	0.997	0.183	0.089
	Topcon-Cirrus-noSVDNA	0.999	0.220	0.155
	Spectralis-Topcon-noSVDNA	0.999	0.218	0.156
	Spectralis-Cirrus-noSVDNA	0.948	0.109	0.178
	Spectralis-Topcon	0.999	0.230	0.164
	Topcon-Cirrus	0.999	0.220	0.165
	Spectralis-Cirrus	0.999	0.235	0.182
	Supervised	0.999	0.232	0.185

Figure 7: Model S+C; Training set Cirrus

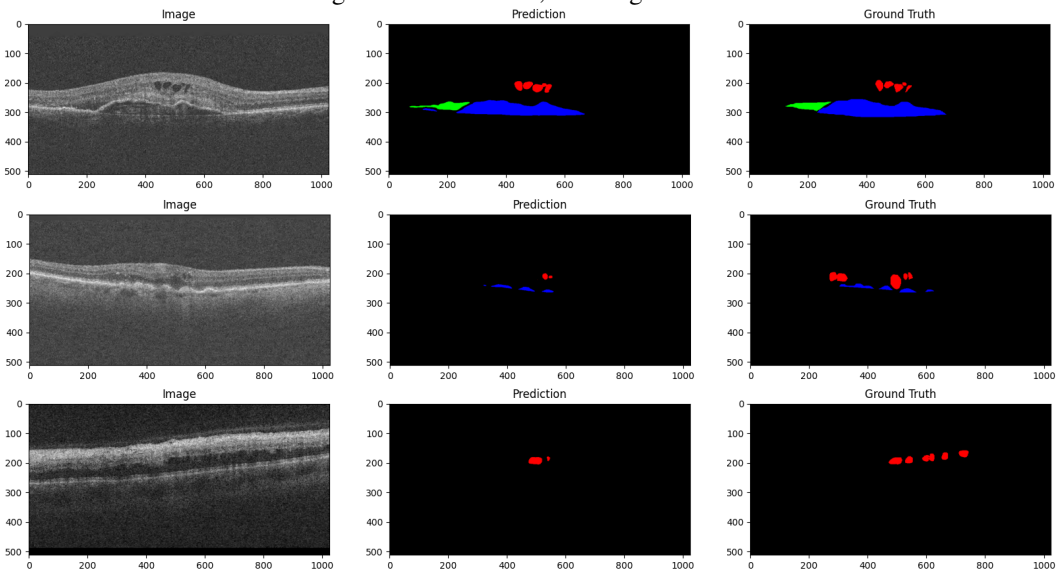


Figure 8: Model S+C; Training set Cirrus; no SVDNA