Reducing the impact of domain shift in deep learning for OCT segmentation using image manipulations

Marc S. Seibel[a], Joshua Niemeijer[b], Marc Rowedder[a], Helge Sudkamp[d], Timo Kepp[c], Gereon Hüttmann[e], and Heinz Handels[a,c]

[a]Institute of Medical Informatics, University of Lübeck, Germany
[b]German Aerospace Center (DLR), Braunschweig, Germany
[c]German Research Center for Artificial Intelligence (DFKI), Lübeck, Germany
[d]Visotec GmbH, Lübeck, Germany
[e]Institute of Biomedical Optics, University of Lübeck, Germany

ABSTRACT

Medical segmentation of optical coherence tomography (OCT) images using deep neural networks (DNNs) has been intensively studied in recent years, but generalization across datasets from different OCT devices is still a considerable challenge. In this work, we focus on the novel self-examination low-cost full-field (SELFF)-OCT, a handheld imaging device for home-monitoring of retinopathies, and the clinically used Spectralis-OCT. Images from both devices exhibit different characteristics, leading to different representations within DNNs and consequently to a reduced segmentation quality when switching between devices. To robustly segment OCT images from an OCT-scanner unseen during training, we alter the appearance of the images using manipulation methods ranging from traditional data augmentation to noise-based methods to learning-based style transfer methods. We evaluate the effect of the manipulation methods with respect to segmentation quality and changes in the feature space of the DNN. Reducing the domain shift with style transfer methods results in a significantly better segmentation of pigment epithelial detachment (PED). Investigations of the feature space show that the segmentation quality of PED is negatively correlated with the distance between training and test distributions. Our methods and results help researchers to choose and evaluate image manipulation methods for developing OCT segmentation models which are robust against domain shifts.

Keywords: OCT, unsupervised domain adaptation, segmentation, image manipulation

1. INTRODUCTION

Age-related macular degeneration (AMD) is a medical condition which results in a loss of vision. AMD can be treated using specific medications which are injected into the eye. This treatment must be repeated based on the condition of the eye. To assess the eye, optical coherence tomography (OCT) has been employed. Home monitoring of the eye has recently been made feasible by the introduction of the self-examination low-cost full-field (SELFF)-OCT[1] which allows the recording of images on a daily basis. As the amount of images increases, physicians need computerized assistance for examining the recorded images of their patients. Biomarkers in these images can be segmented by employing deep neural networks (DNNs), but training them for the domain of SELFF-OCT is challenging, since a lack of pixel-wise annotations exist.[2] One approach is to pretrain DNNs on larger datasets which were recorded using other OCT devices. Problematically, the different image characteristics of the pretraining and SELFF-OCT dataset lead to a drop of segmentation quality.

This phenomenon is called domain shift and an actively studied topic.[3–6] We study the domain shift problem under the assumption that labels are only given for a training (source) domain but not the test (target) domain. Domain adaptation can be categorized depending on the adaptation spaces, which are the input space, the feature space, and the output space of a DNN. Here, we study how manipulations in the input space affect the segmentation quality. Segmentation metrics alone do not explain why a given manipulation method changes the segmentation quality. To provide further insight, we measure the representation shifts in the feature space using the univariate Wasserstein distance, as proposed by Stacke et al. 7.

Corresponding author: Marc S. Seibel
E-mail: marc.seibel@uni-luebeck.de

Medical Imaging 2024: Computer-Aided Diagnosis, edited by Weijie Chen,
Susan M. Astley, Proc. of SPIE Vol. 12927, 1292719 · © 2024
SPIE · 1605-7422 · doi: 10.1117/12.3008702

2. RELATED WORK

Unsupervised Domain Adaptation: Unsupervised Domain Adaptation (UDA) involves the adaptation from a labeled source domain distribution to an unlabeled target domain distribution. This sets it apart from e.g. tasks like domain generalization[8] where the target domain distribution is unknown or semi supervised domain adaptation[9] where a small subset of the target domain is labeled. UDA approaches can be divided in three categories:[3] Adaptation in the input, the feature and the output space of a neural network. Feature space adaptation is e.g. done by alignment of the feature space distributions of source and target domain by clustering as e.g. in Niemeijer et al.[4] or through adversarial training as in Hoffman et al.,[10] Li et al.[11] or Wang et al.[12] Output space adaptation usually consists of self training on the target domain as e.g. applied in Zheng et al.[13] or adversarial training on the predicted outputs as in Tsai et al.[14] But since we are studying the manipulation of the input space, the input space adaptation is the most interesting to our work. The input space adaptation is mostly performed by computing a style transfer between the source and the target domain. This is often done by utilizing CycleGANs as in e.g. Seebock et al.[15] or Romo et al.[16] CycleGANs however have the problem that in failure cases they not only alter the style of an image but also the content. We focused on a non-adversarial approach that is based on adaptive instance normalisation as introduced in Huang et al.[17] and further refined in Liu et al.[18] Such approaches use an auto encoder. The content image that should be style transformed, is passed through it, as well as a style image. The channel-wise mean and variance statistics of the content image feature space is hereby exchanged with that of the style image. The style image is hereby chosen from the target domain and the content image from the source domain.

Measuring domain shift: The problem of measuring domain shifts can be seen as a subset of approaches which are intended to perform out-of-distribution detection, in which the classification question is posed whether a data point belongs to the source domain or to a target domain.[19] These methods decide at the output stage of a neural network whether a sample belongs to the source domain using entropy measures.[20–22] Likewise, You et al.[23] assess the transferability of models for a given dataset. More recently, the L2-distance has been used to measure the distance of source and target samples in the feature space by Sun et al.[24] Closest to our use case, Stacke et al. argue for the Wasserstein distance to quantify the effect of domain shifts and image manipulations.[7]

3. METHODS AND MATERIALS

In this chapter, we first describe our experiment setup with respect to the data that is used to investigate the effects of the image manipulation methods. We then introduce the methods for image manipulation and the approach for measuring the domain shift.

3.1 Dataset preparation

SELFF-OCT/ Spectralis-OCT: The patients in this dataset were diagnosed with neovascular AMD and showed at least at one eye signs of a medical condition. The dataset contains a total of 45 patients, from which we excluded seven patients due to low image quality and missing correspondence in either the Spectralis-OCT or the SELFF-OCT dataset split. All images are accompanied by a pixel-wise segmentation of subretinal fluids (SRF), intraretinal fluids (IRF), and pigment epithelial detachment (PED), the retina, and the background. A complete description of the dataset can be found in.[25] To learn further topographic information, we refined the annotations by separating the background annotation into vitreous humor and choroid (areas above and below retina). The IRF class was mapped to the retina class, because annotators expressed concern about the reliability of the IRF annotation. Thus, we have $C = 5$ classes for training the segmentation network. To account for the different resolutions of the B-scans, the B-scans of the Spectralis device are resampled to the same resolution as the B-scans of the SELFF-OCT scanner. In total, 3124 and 2048 B-scans for the Spectralis and SELFF-OCT device are used. Note that for some patients, we included only one eye because the excluded eye was missing for one of the OCT devices. For training, we randomly take patches of size 256×512 pixels.

RETOUCH: The RETOUCH dataset has been described in Ref.[26] It contains pixel-wise segmentation for the IRF, SRF, and PED. No segmentation for the retina exists, thus we can not separate the background annotation into vitreous humor and choroid, as we did for the SELFF-OCT dataset. Here, we use a subset of the dataset, containing only the Spectralis-OCT and Topcon-OCT (T-1000 and T-2000) images. To account for the different resolutions of the B-scans, we resampled all B-scans to a lateral resolution of 11.72 × 3.5 µm. For training, we randomly take patches of size 512 × 512 pixels.

3.2 Manipulation methods

We compare four methods for manipulating images. The manipulations are intended to reduce the domain shift between the images from different devices. As a baseline, we use the original images. The first method is supposed to increase the diversity of the dataset and thereby the robustness of the segmentation model. It consists of a traditional augmentation pipeline, which randomly applies gamma transformation, additive intensity shifts, and histogram shifts. The second method reduces the noise of the images using the structured Noise2Void (N2V) algorithm.[27] The third method and fourth method aim at increasing the similarity of the images from both devices. The singular value decomposition noise adaptation (SVDNA) algorithm translates the noise and the histograms between images from different devices.[28] The adaptive attention normalization (AdaAttN) algorithm[18] is a DNN-based method to translate the style from one image to another and thereby increase their similarity. Figure 1 shows the effect of the manipulation methods.

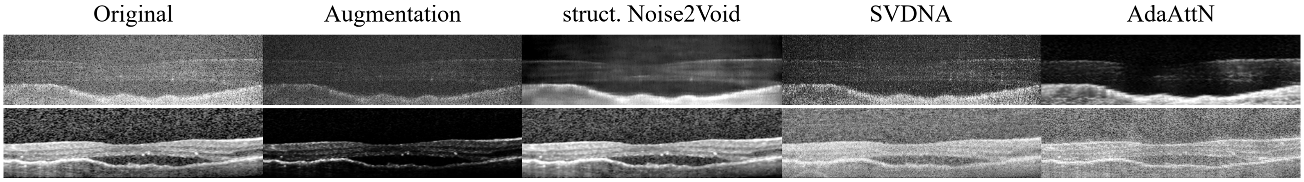

Figure 1. Visualization of the manipulation methods. The first row displays images from the SELFF-OCT domain, while the second row shows images from the Spectralis-OCT domain along with the applied manipulations. The same patient is shown for both domains.

3.3 Measuring domain shift

Oriented on the approach of Stacke et al. 7, we use univariate Wasserstein distance for measuring the feature distance between different datasets mapped to a neural network's feature space. The Wasserstein distance between two distributions ρ_s, ρ_t is given as,

$$W(\rho_s, \rho_t) = \inf_{\pi \in \Gamma(\rho_s, \rho_t)} \int_{\mathbb{R} \times \mathbb{R}} \|x - y\| d\pi(x, y). \tag{1}$$

To obtain the distributions ρ, we sample $n = 30$ feature vectors z_i from the embedded OCT image at the positions i in the feature space. We choose the positions based on the ground truth segmentation masks. That way, we can compare the feature distributions that belong to the same class. Additionally, for the SELFF-OCT/Spectralis-OCT dataset we compare the feature representations in a patient-paired and eye-paired manner, i.e. we compute the Wasserstein distance $W(\rho_s^{(p,e,c)}, \rho_t^{(p,e,c)})$ where p, e, c stands for the same patient, eye, and class. The obtained distance values are averaged over all available subjects and eyes. The RETOUCH dataset does not have paired subjects in the source and target datasets. Thus, we aggregate the feature vectors from all images and subjects to form the distributions and subsequently compute the distance.

3.4 Network architecture and training settings

For robust segmentation, we chose the based HRDA architecture of,[29] because Transformer based architectures showed stronger robustness against domain shifts.[6] As encoder, we chose the DAFormer[6] and the HRDA decoder was chosen because it addresses the trade-off between a manageable GPU memory footprint and high resolution images. Batch sizes were three and four for the SELFF-OCT/Spectralis-OCT and RETOUCH dataset. We train the model for 20,000 iterations using only training data from the source domain and create model snapshots after every 1000 iteration using the target validation dataset. The weights associated with the best snapshot are then selected for final testing on the test dataset.

4. RESULTS

Evaluation design: To evaluate the employed manipulation methods, we split the data according to their sensor domains (source and target split). For each domain, we created development and test sets using five-fold cross validation. The development sets were again split into training and validation sets using an 80-20 ratio. Patient-related data leakage was prevented by ensuring that patients from the training set were not included in the test set. We trained the segmentation model for each source domain separately (Spectralis-OCT, SELFF-OCT). The validation set from the target domain was used to select the best model during the training process. Final evaluation was performed using the test set. Training and evaluation were repeated for each fold and image manipulation method. Segmentation quality and representation shift was measured for each fold and class using the Dice score and the Wasserstein distance. Based on the five folds, we calculated p-values for the Dice scores using the paired t-test. Normality was tested using the Shapiro-Wilk test.

Evaluation SELFF-OCT/ Spectralis-OCT: In Fig. 2, we show the effect of the manipulation methods on the segmentation quality of the PED. When training and test data comes from the same device, training with manipulation methods did not improve over traditional training in which only data augmentations (or no augmentations) are used (Spectralis-OCT: $p = 0.70$, SELFF-OCT: $p = 0.20$). This is as expected, since the style transfer translates the source data to the target domain and thereby increases the discrepancy between the training and test data. For the domain adaptation setting, we find that applying AdaAttN during training improves the Spectralis-OCT to SELFF-OCT segmentation quality from 24.9% to 29.7% ($p < 0.05$), while testing with manipulated images decreased the performance. For the SELFF-OCT to Spectralis-OCT shift, testing with AdaAttN increases the performance from 21.8% to 29.2% ($p < 0.05$). For other classes besides PED, the best performing manipulation varies, as detailed in Fig. 6 in the appendix. Subsequently, we investigated the effect of the manipulation methods on the feature representation within the network. We show in Fig. 3 the t-SNE embedding at an early and the pre-logit layer in the network. After the first layer of the network, individual clusters for the manipulation methods can be seen. At the pre-logit layer, the domains are clearly separable. Using the Wasserstein distance in the 256 dimensional feature space of the pre-logit layer, we find a correlation

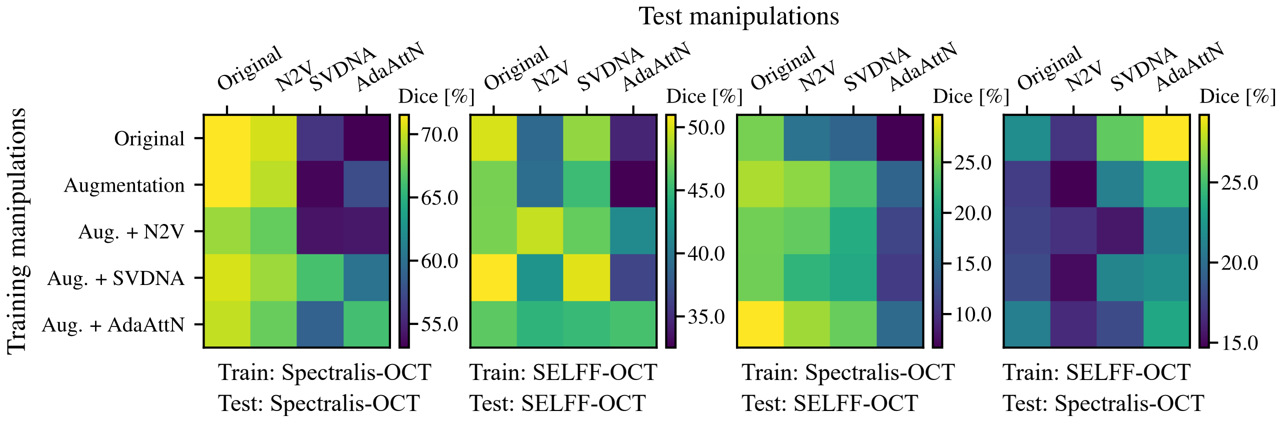

Figure 2. The heatmaps show the Dice score for the segmentation of the PED for different combinations of manipulations applied at train and test time. The first and second heatmaps show the effect of the manipulations if the network is tested in the source-only setting, i.e. the training and test domain is the same. The third and fourth heatmap show the effect of manipulation methods in the domain adaptation setting where the test domain differs from the training domain. For example, the third heatmap shows that training with AdaAttN translated Spectralis-OCT images and testing with the original SELFF-OCT images increases the segmentation quality compared to the baseline.

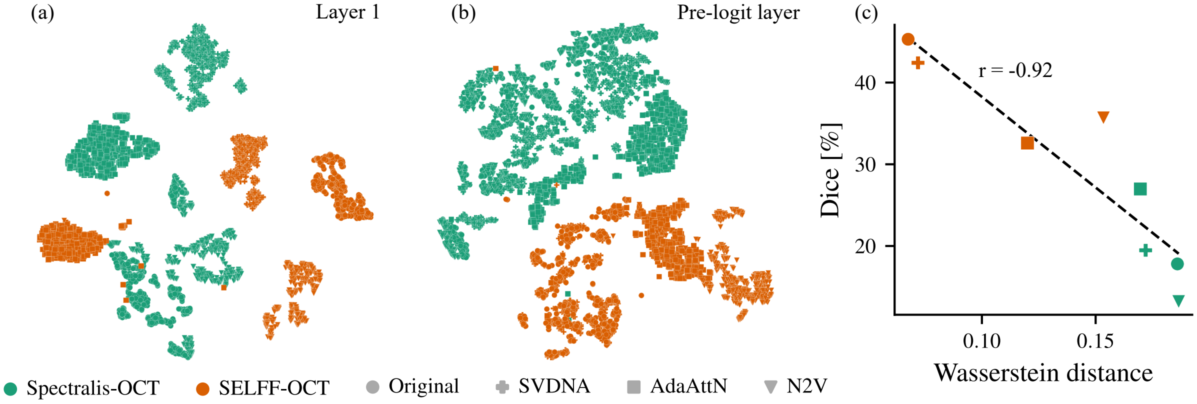

Figure 3. Visualization of the effect of manipulation methods on the feature space. We examine a DNN which was trained with the original SELFF-OCT data. Plot (a) and (b) show the t-SNE embeddings of feature vectors at the first and at the pre-logit layer of the DNN. Marker color and form must be combined. E.g. 🟧 refers to an SELFF-OCT image whose style was translated towards the Spectralis-OCT using AdaAttN. Plot (a) shows that the manipulation methods help to mix shallow image characteristics of the domains. Deeper image characteristics, as displayed in the second plot, are roughly linearly separable. In plot (c), the Dice score is plotted against the Wasserstein distance measured in the 256 dimensional space at the pre-logit layer of the decoder. Best viewed with zoom.

of −0.92 with the Dice score. In other words, a reduction of the distance is related with a better segmentation. The corresponding data can be seen in the scatter plot in Fig. 3 (c).

Evaluation RETOUCH Topcon-OCT/ Spectralis-OCT: To obtain additional evidence, we conducted the same evaluation on the RETOUCH dataset. The heatmaps in Fig. 4 show that the SVDNA method provides the best image manipulations for the Spectralis-OCT trained models in the source setting (column 1) and in the domain shift setting (column 3). For the Topcon-OCT trained models, a combination of data augmentation and denoising at test time performs best in the source-only and domain shift setting (columns 2 and 4).

In Fig. 5, we study the feature space for the Topcon-OCT to Spectralis-OCT domain shift. The t-SNE plots show especially in the first layer a larger homogeneity for both embedded dataset. After the pre-logit layer, the

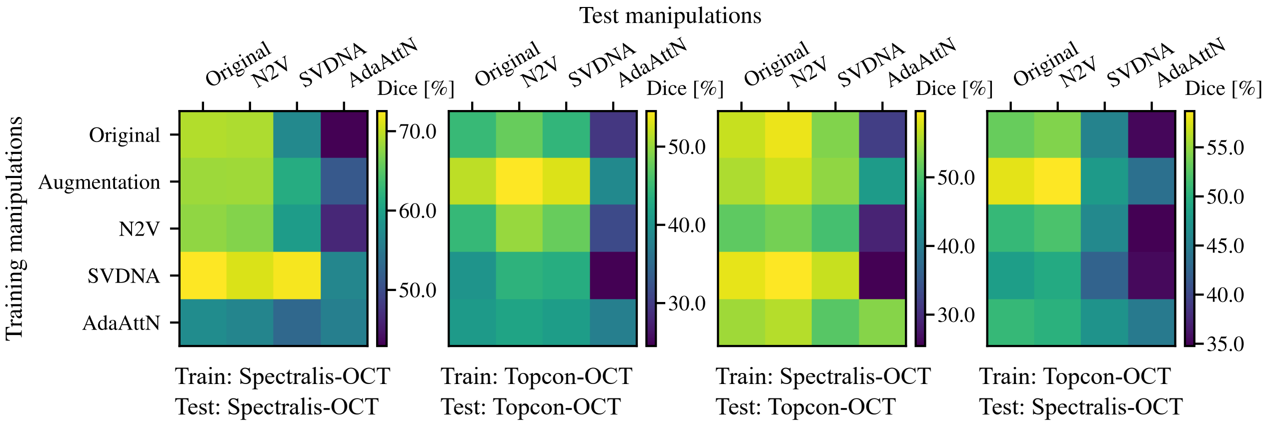

Figure 4. The heatmaps show the Dice score for the segmentation of the PED for different combinations of manipulations applied at train and test time. The first and second heatmaps show the effect of the manipulations if the network is tested in the source-only setting, i.e. the training and test domain is the same. The third and fourth heatmap show the effect of manipulation methods in the domain adaptation setting where the test domain differs from the training domain.

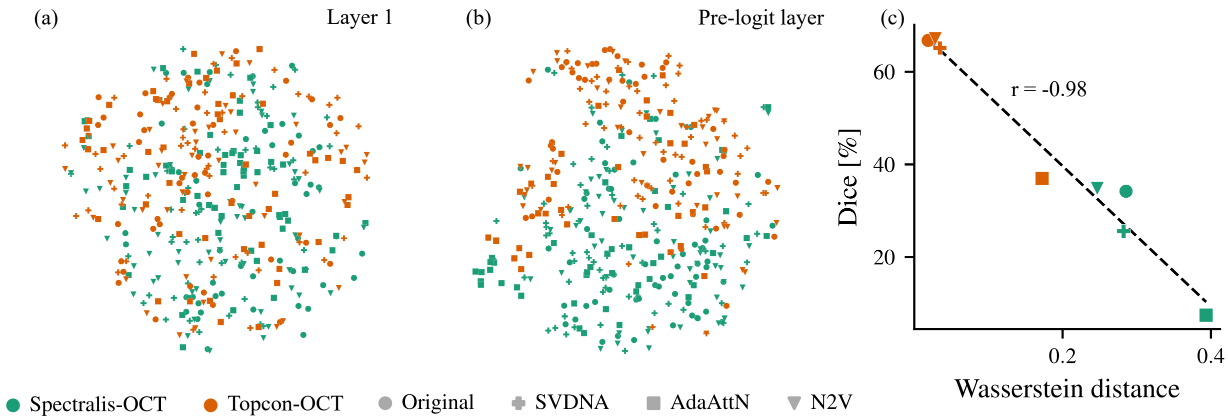

Figure 5. Visualization of the effect of manipulation methods on the feature space. The examined DNN was trained with the original Topcon-OCT images. Plot (a) shows that the domains are mixed at layer 1, while plot (b) shows that the domains are more distinguishable at the pre-logit layer. Plot (c) shows that feature distance and segmentation performance are strongly correlated.

embeddings of both domains become better distinguishable. Finally, we see a strong correlation for a relationship between the Wasserstein distance and the Dice score.

5. DISCUSSION AND CONCLUSION

In our experiments, we compared image manipulation methods with the aim to improve the segmentation quality in the presence of domain shifts. To give insight beyond segmentation metrics, we studied the feature space qualitatively using t-SNE plots and quantitatively by measuring the domain shift using the Wasserstein distance. Precisely, our results were conducted to help us answer three questions related to finding optimal image manipulation methods:

1. Does a given manipulation method improve the segmentation quality on the target domain?
2. Does a given manipulation method improve the segmentation quality on the source domain?
3. Is the effect of manipulation methods measurable in the feature space?

We first found that style transfer-based methods can be used to increase the segmentation quality in the target domain at training and at test time. When comparing the results of the SELFF-OCT dataset with the RETOUCH dataset, we saw that for the first dataset AdaAttN improved the segmentation quality while for the second dataset SVDNA led to better results. This shows that there is no go-to method for improving the segmentation quality on a target domain. In particular, the direction of the domain shift and the class to be segmented seem to be factors which should be considered when choosing an image manipulation method. Secondly, we found that the SVDNA method improved slightly the performance for in-distribution segmentation. This effect was consistent for Spectralis-OCT data from both datasets. Thirdly, our study of the feature space (a) showed that domain shifts become more visible in the deeper layers of the network, and (b) confirmed the observation of Stacke et al.[7] that if an image manipulation method decreased the distance between features from the source and target domain, then the segmentation quality improved. As limitations of our study, we need to stress that our results do not provide evidence in favor of one or another image manipulation method. That is, our results do not allow us to predict which image manipulation method will work particular well for a domain shift outside this study. Further, the training of the AdaAttN model provides another source of uncertainty, since we relied on visual inspection to determine the quality of the style transfer. In conclusion, our work presents data for the effectiveness of image manipulation methods and shows that the effect of these methods can also be seen and measured in the feature space. Future work could be focused on translating the observed relationship between the Wasserstein distance and the Dice score into a tool which predicts the segmentation quality and thereby is useful for selecting and adapting image manipulation methods.

REFERENCES

[1] Sudkamp, H., Koch, P., Spahr, H., Hillmann, D., Franke, G., Münst, M., Reinholz, F., Birngruber, R., and Hüttmann, G., "In-vivo retinal imaging with off-axis full-field time-domain optical coherence tomography," *Optics Letters* **41**, 4987–4990 (Nov 2016).

[2] Kepp, T., Andresen, J., Sudkamp, H., von der Burchard, C., Roider, J., Hüttmann, G., Ehrhardt, J., and Handels, H., "Epistemic and aleatoric uncertainty estimation for ped, segmentation in home oct images," in [*German Workshop on Medical Image Computing,*], Maier-Hein, K., Deserno, T. M., Handels, H., Maier, A., Palm, C., and Tolxdorff, T., eds., *Informatik aktuell*, 32–37, Springer Fachmedien (Mar. 2022).

[3] Schwonberg, M., Niemeijer, J., Termöhlen, J.-A., Schäfer, J. P., Schmidt, N. M., Gottschalk, H., and Fingscheidt, T., "Survey on unsupervised domain adaptation for semantic segmentation for visual perception in automated driving," *IEEE Access* **11**, 54296–54336 (May 2023).

[4] Niemeijer, J., Ehrhardt, J., Kepp, T., Schäfer, J. P., and Handels, H., "Overcoming the sensor delta for semantic segmentation in OCT images," in [*Medical Imaging 2023: Computer-Aided Diagnosis*], Iftekharuddin, K. M. and Chen, W., eds., 34, SPIE, San Diego, United States (Apr. 2023).

[5] Niemeijer, J. and Schäfer, J. P., "Domain adaptation and generalization: A low-complexity approach," in [*Proceedings of The 6th Conference on Robot Learning*], Liu, K., Kulic, D., and Ichnowski, J., eds., *Proceedings of Machine Learning Research* **205**, 1081–1091, PMLR (14–18 Dec 2023).

[6] Hoyer, L., Dai, D., and Van Gool, L., "Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*], 9924–9935 (June 2022).

[7] Stacke, K., Eilertsen, G., Unger, J., and Lundström, C., "Measuring Domain Shift for Deep Learning in Histopathology," **25**, 325–336 (Feb 2021).

[8] Niemeijer, J., Schwonberg, M., Termöhlen, J.-A., Schmidt, N. M., and Fingscheidt, T., "Generalization by adaptation: Diffusion-based domain extension for domain-generalized semantic segmentation," in [*Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*], 2830–2840 (January 2024).

[9] Mütze, A., Rottmann, M., and Gottschalk, H., "Semi-supervised domain adaptation with cyclegan guided by a downstream task loss," *arXiv preprint arXiv:2208.08815* (2022).

[10] Hoffman, J., Wang, D., Yu, F., and Darrell, T., "Fcns in the wild: Pixel-level adversarial and constraint-based adaptation," *arXiv preprint arXiv:1612.02649* (2016).

[11] Li, T., Huang, K., Zhang, Y., Li, M., Zhang, W., and Chen, Q., "Multi-stage domain adaptation for subretinal fluid classification in cross-device oct images," in [*Pattern Recognition*], Wallraven, C., Liu, Q., and Nagahara, H., eds., 474–487, Springer International Publishing, Cham (2022).

[12] Wang, J., Chen, Y., Li, W., Kong, W., He, Y., Jiang, C., and Shi, G., "Domain adaptation model for retinopathy detection from cross-domain oct images," in [*Proceedings of the Third Conference on Medical Imaging with Deep Learning*], Arbel, T., Ben Ayed, I., de Bruijne, M., Descoteaux, M., Lombaert, H., and Pal, C., eds., *Proceedings of Machine Learning Research* **121**, 795–810, PMLR (06–08 Jul 2020).

[13] Zheng, Z. and Yang, Y., "Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation," *International Journal of Computer Vision* , 1–15 (2021).

[14] Tsai, Y.-H., Hung, W.-C., Schulter, S., Sohn, K., Yang, M.-H., and Chandraker, M., "Learning to Adapt Structured Output Space for Semantic Segmentation," in [*Proc. of CVPR*], 7472–7481 (June 2018).

[15] Seeböck, P., Romo-Bucheli, D., Waldstein, S., Bogunovic, H., Orlando, J. I., Gerendas, B. S., Langs, G., and Schmidt-Erfurth, U., "Using cyclegans for effectively reducing image variability across oct devices and improving retinal fluid segmentation," in [*2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*], 605–609, IEEE (2019).

[16] Romo-Bucheli, D., Seeböck, P., Orlando, J. I., Gerendas, B. S., Waldstein, S. M., Schmidt-Erfurth, U., and Bogunović, H., "Reducing image variability across oct devices with unsupervised unpaired learning for improved segmentation of retina," *Biomed. Opt. Express* **11**, 346–363 (Jan 2020).

[17] Huang, X. and Belongie, S., "Arbitrary style transfer in real-time with adaptive instance normalization," in [*Proceedings of the IEEE International Conference on Computer Vision (ICCV)*], (Oct 2017).

[18] Liu, S., Lin, T., He, D., Li, F., Wang, M., Li, X., Sun, Z., Li, Q., and Ding, E., "AdaAttN: Revisit Attention Mechanism in Arbitrary Neural Style Transfer," in [*2021 IEEE/CVF International Conference on Computer Vision (ICCV)*], 6629–6638 (2021).

[19] Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F., "Analysis of Representations for Domain Adaptation," in [*Advances in Neural Information Processing Systems*], Schölkopf, B., Platt, J., and Hoffman, T., eds., **19**, MIT Press (2006).

[20] Mehrtash, A., Wells, W. M., Tempany, C. M., Abolmaesumi, P., and Kapur, T., "Confidence Calibration and Predictive Uncertainty Estimation for Deep Medical Image Segmentation," *IEEE Transactions on Medical Imaging* **39**, 3868–3878 (Dec. 2020). Conference Name: IEEE Transactions on Medical Imaging.

[21] Araújo, T., Aresta, G., Schmidt-Erfurth, U., and Bogunović, H., "Few-shot out-of-distribution detection for automated screening in retinal OCT images using deep learning," *Scientific Reports* **13**, 16231 (Sept. 2023). Number: 1 Publisher: Nature Publishing Group.

[22] Garg, S., Balakrishnan, S., Lipton, Z. C., Neyshabur, B., and Sedghi, H., "Leveraging unlabeled data to predict out-of-distribution performance," *arXiv preprint arXiv:2201.04234* (2022).

[23] You, K., Liu, Y., Wang, J., and Long, M., "Logme: Practical assessment of pre-trained models for transfer learning," in [*International Conference on Machine Learning*], 12133–12143, PMLR (2021).

[24] Sun, Y., Tzeng, E., Darrell, T., and Efros, A. A., "Unsupervised domain adaptation through self-supervision," *CoRR* **abs/1909.11825** (2019).

[25] von der Burchard, C., Sudkamp, H., Tode, J., Ehlken, C., Purtskhvanidze, K., Moltmann, M., Heimes, B., Koch, P., Münst, M., Endt, M. v., Kepp, T., Theisen-Kunde, D., König, I., Hüttmann, G., and Roider, J., "Self-Examination Low-Cost Full-Field Optical Coherence Tomography (SELFF-OCT) for neovascular age-related macular degeneration: a cross-sectional diagnostic accuracy study," *BMJ Open* **12**, e055082 (June 2022). Publisher: British Medical Journal Publishing Group Section: Ophthalmology.

[26] Bogunovic, H., Venhuizen, F., Klimscha, S., Apostolopoulos, S., Bab-Hadiashar, A., Bagci, U., Beg, M. F., Bekalo, L., Chen, Q., Ciller, C., Gopinath, K., Gostar, A. K., Jeon, K., Ji, Z., Kang, S. H., Koozekanani, D. D., Lu, D., Morley, D., Parhi, K. K., Park, H. S., Rashno, A., Sarunic, M., Shaikh, S., Sivaswamy, J., Tennakoon, R., Yadav, S., De Zanet, S., Waldstein, S. M., Gerendas, B. S., Klaver, C., Sanchez, C. I., and Schmidt-Erfurth, U., "RETOUCH: The Retinal OCT Fluid Detection and Segmentation Benchmark and Challenge," **38**(8), 1858–1874 (2019).

[27] Broaddus, C., Krull, A., Weigert, M., Schmidt, U., and Myers, G., "Removing Structured Noise with Self-Supervised Blind-Spot Networks," in [*2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*], 159–163 (Apr 2020).

[28] Koch, V., Holmberg, O., Spitzer, H., Schiefelbein, J., Asani, B., Hafner, M., and Theis, F. J., "Noise transfer for unsupervised domain adaptation of retinal OCT images," **13432**, 699–708 (Oct 2022).

[29] Hoyer, L., Dai, D., and Van Gool, L., "HRDA: Context-Aware High-Resolution Domain-Adaptive Semantic Segmentation," in [*Computer Vision – ECCV 2022*], Avidan, S., Brostow, G., Cissé, M., Farinella, G. M., and Hassner, T., eds., **13690**, 372–391, Springer Nature Switzerland, Cham (Oct. 2022). Series Title: Lecture Notes in Computer Science.

APPENDIX A. PERFORMANCE FOR TRAIN- AND TEST-MANIPULATIONS

In the main text, we presented only the segmentation quality for the PED. Notably, the effectiveness of the image manipulations seems to depend on the class to be segmented and on the direction of the domain shift, as is shown in Figures 6 and 7. While the style transfer using the AdaAttN method presented itself effective for improving the segmentation performance by turning Spectralis-OCT into SELFF-OCT images, no evidence for its effectiveness was found for the RETOUCH dataset. On the other hand, the SVDNA method showed itself beneficial for in- and out of distribution generalization when training DNNs with Spectralis-OCT data. The N2V method did rarely improve the segmentation quality for the SELFF-OCT/ Spectralis-OCT dataset, but N2V was useful for segmenting Topcon images.

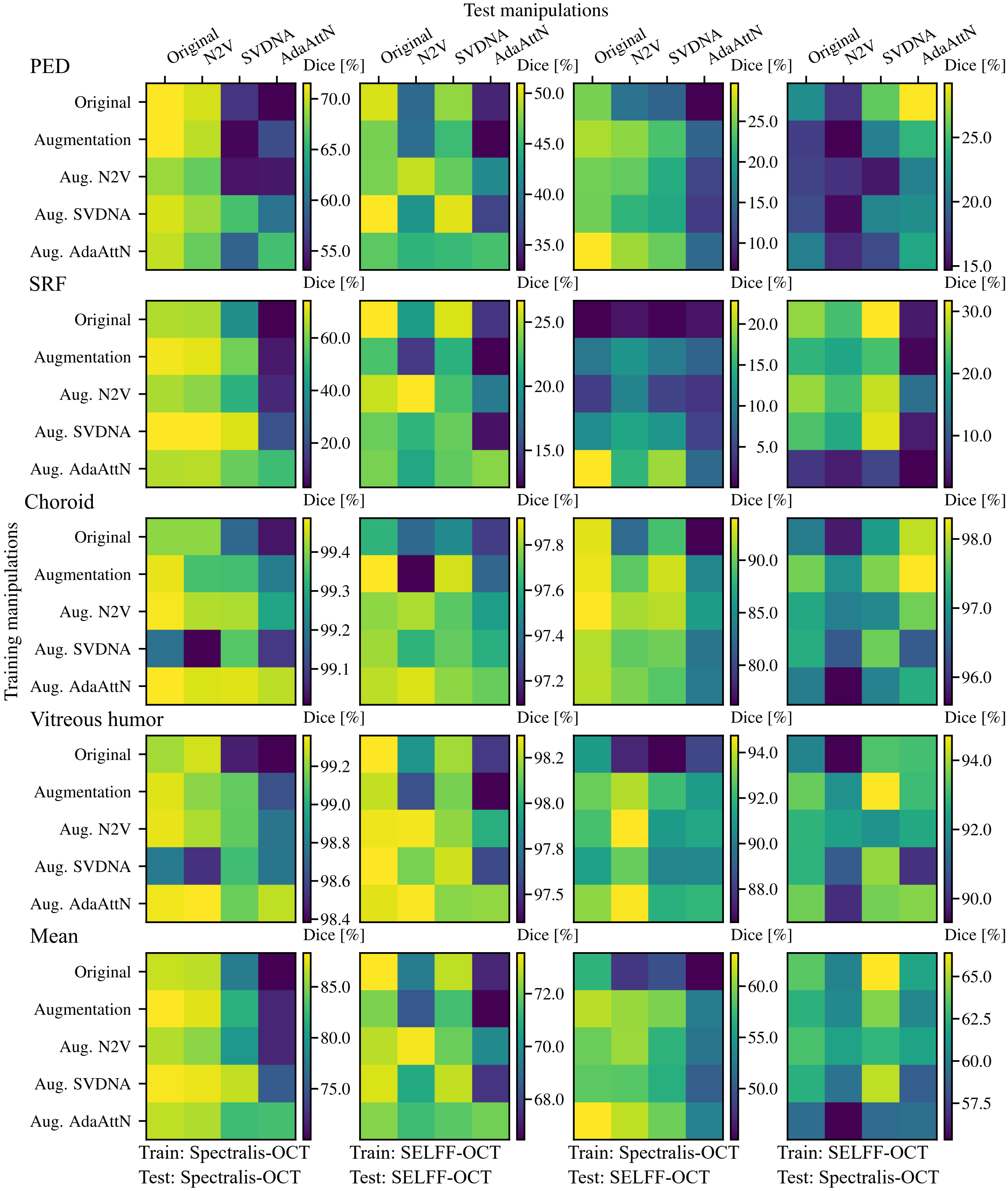

Figure 6. Dice scores for the SELFF-OCT/ Spectralis-OCT dataset w.r.t. different data manipulations at training and test time. From top to bottom the Dice score per class (PED, SRF, choroid, vitreous humor), and the mean is displayed.

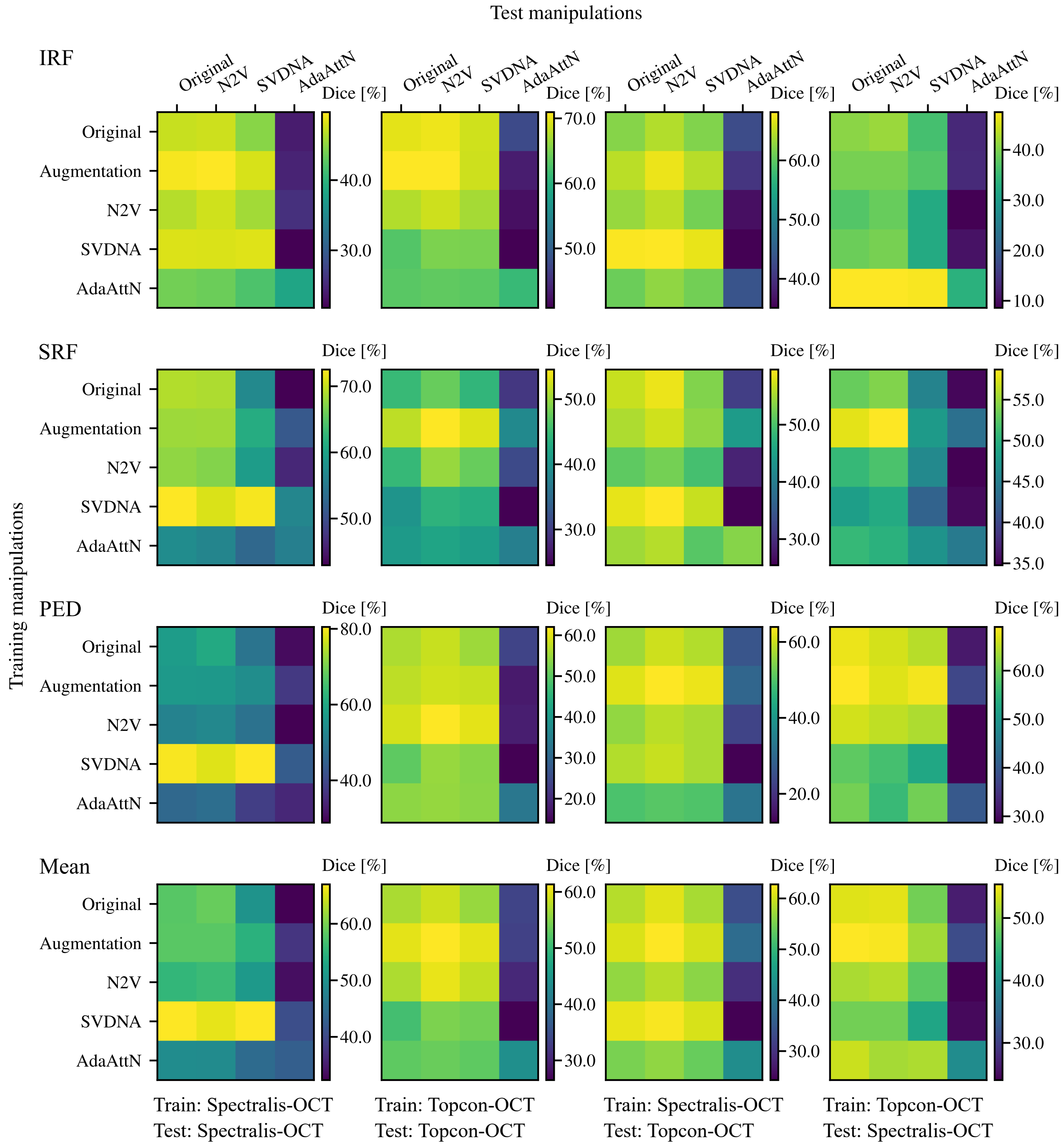

Figure 7. Dice scores for the RETOUCH dataset with respect to different data manipulations at training and test time. The figure shows from top to bottom the Dice score per class (IRF, SRF, PED) and finally the mean Dice.