Self-supervised Learning via Inter-modal Reconstruction and Feature Projection Networks for Label-Efficient 3D-to-2D Segmentation

José Morano[1](✉), Guilherme Aresta[1], Dmitrii Lachinov[1], Julia Mai[2], Ursula Schmidt-Erfurth[2], and Hrvoje Bogunović[1,2]

[1] Christian Doppler Laboratory for Artificial Intelligence in Retina, Department of Ophthalmology and Optometry, Medical University of Vienna, Vienna, Austria
{jose.moranosanchez,hrvoje.bogunovic}@meduniwien.ac.at

[2] Lab for Ophthalmic Image Analysis, Department of Ophthalmology and Optometry, Medical University of Vienna, Vienna, Austria

Abstract. Deep learning has become a valuable tool for the automation of certain medical image segmentation tasks, significantly relieving the workload of medical specialists. Some of these tasks require segmentation to be performed on a subset of the input dimensions, the most common case being 3D→2D. However, the performance of existing methods is strongly conditioned by the amount of labeled data available, as there is currently no data efficient method, e.g. transfer learning, that has been validated on these tasks. In this work, we propose a novel convolutional neural network (CNN) and self-supervised learning (SSL) method for label-efficient 3D→2D segmentation. The CNN is composed of a 3D encoder and a 2D decoder connected by novel 3D→2D blocks. The SSL method consists of reconstructing image pairs of modalities with different dimensionality. The approach has been validated in two tasks with clinical relevance: the en-face segmentation of geographic atrophy and reticular pseudodrusen in optical coherence tomography. Results on different datasets demonstrate that the proposed CNN significantly improves the state of the art in scenarios with limited labeled data by up to 8% in Dice score. Moreover, the proposed SSL method allows further improvement of this performance by up to 23%, and we show that the SSL is beneficial regardless of the network architecture.

Keywords: image segmentation · self-supervised learning · OCT · retina

1 Introduction

Deep learning can significantly reduce the workload of medical specialists during image segmentation tasks, which are essential for patient diagnosis and

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43901-8_56.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14223, pp. 589–599, 2023.
https://doi.org/10.1007/978-3-031-43901-8_56

follow-up management [8,14,16]. For most tasks, segmentation masks have the same dimensionality as the input. However, there are some tasks for which segmentation has to be performed in a subset of the dimensions of the data, e.g. 3D→2D [13,21]. This occurs, for example, for the segmentation of geographic atrophy (GA) in optical coherence tomography (OCT), where the segmentation is performed on the OCT projection. In recent years, several methods have been proposed for this type of tasks [9,11–13]. Li *et al.* [11] proposed an image projection network (IPN) that reduces the features to the target dimensionality using unidirectional pooling layers in the encoder. However, IPN follows a patch-based approach with fixed patch size, which prevents its direct application to full 3D volumes of varying size. Also, it does not have skip connections, which have proven to be highly useful for accurate segmentation. Later, Lachinov *et al.* [9] proposed a U-Net-like convolutional neural network (CNN) for 3D→2D segmentation that overcomes the limitations of IPN, which were also later overcome by the second version of IPN (IPNv2) [12]. However, they still require a large amount of labeled data to provide adequate performance. In addition, there are works that explore the use of CNNs for 3D→2D regression, where Seeböck *et al.* [20] proposed ReSensNet, a novel CNN based on Residual 3D U-Net [10], with a 3D encoder and a 2D decoder connected by 3D→2D blocks. However, ReSensNet only works at concrete input resolutions, and it is applied pixel-wise.

In general, one of the issues of these and other deep learning segmentation methods is that their performance strongly depends on the amount of annotated data [22], which hinders their deployment to real-world medical image analysis settings. Transfer learning from ImageNet is the standard approach to mitigate this issue [22]. However, specifically for segmentation, ImageNet pre-training has shown minimal performance gains [5,17], partially because it can only be performed on the encoder part of the very common encoder-decoder architectures.

A possible alternative is to pre-train the models using a self-supervised learning (SSL) paradigm [1,2,4,6,7,15,18]. However, only some of these approaches have the potential to be applied for 3D→2D segmentation, as many of them, such as image denoising [2], require input and output images to have the same dimensionality. Among the suitable approaches, multi-modal reconstruction pre-training (MMRP) shows great potential in multi-modal scenarios [7]. In this approach, models are trained to reconstruct pairs of images from different modalities, learning relevant patterns in the data without requiring manual annotations. MMRP, however, has only been proven useful for localizing non-pathological structures on 2D color fundus photography, using fluorescein angiography as the modality to reconstruct. Moreover, image pairs of these modalities have to be registered using a separate method.

Contributions. In this work, we propose a novel approach for label-efficient 3D→2D segmentation. In particular, our contributions are as follows: (1) As an alternative to state-of-the-art network architectures, we propose a 3D→2D segmentation CNN based on ReSensNet [20] that has a 3D encoder and a 2D decoder connected by novel 3D→2D projective blocks. (2) We propose a novel SSL strategy for 3D→2D models based on the reconstruction of modalities of

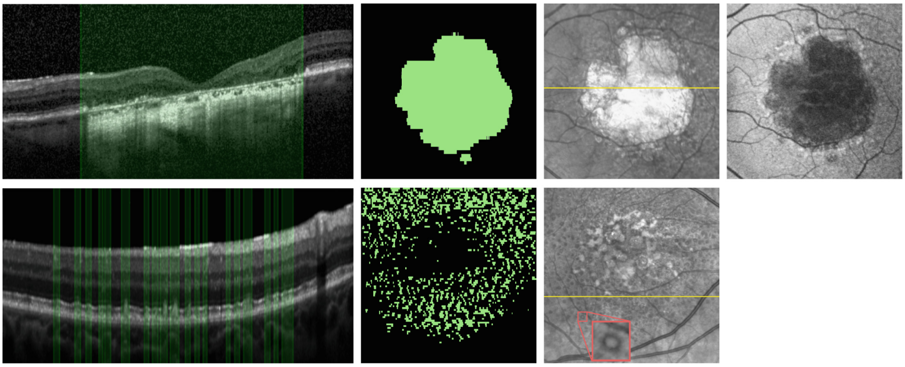

Fig. 1. From left to right: OCT slice (B-scan) with the corresponding ground truth annotations overlaid in green, ground truth, SLO with the location of the B-scan indicated in yellow and a zoom-in view in red, and FAF. Top: GA. Bottom: RPD. (Color figure online)

different dimensionality, and show that it significantly improves the performance of the models in the target segmentation tasks. This is the first data efficient method proposed for 3D→2D models and the first work exploring 3D→2D reconstruction. (3) Lastly, the performed experiments deepen the understanding of the proposed SSL paradigm, by exploring different settings with different image modalities. The proposed approach was validated on two clinically-relevant tasks: the en-face segmentation of GA and reticular pseudodrusen (RPD) in retinal OCT. The results demonstrate that the proposed approach clearly outperforms the state of the art in scenarios with scarce labeled data. Our code is publicly available on GitHub[1].

Clinical Background. *Geographic atrophy* (GA) is an advanced form of age-related macular degeneration (AMD) that corresponds to a progressive loss of retinal photoreceptors and leads to irreversible visual impairment. GA is typically assessed with OCT and/or fundus autofluorescence (FAF) imaging modalities [3,24]. In OCT, it is characterized by the loss of retinal pigment epithelium (RPE) tissue, accompanied by the contrast enhancement of the signal below the retina, and in FAF, by the loss of RPE autofluorescence [19] (see Fig. 1). In both cases, GA lesion is delineated as a 2D en-face area. Also, GA frequently appears brighter than the surrounding areas on scanning laser ophthalmoscopy (SLO) images due to its higher reflectance.

Reticular pseudodrusen (RPD) are accumulations of extracellular material that commonly occur in association with AMD. In OCT scans, these lesions are shown as granular hyperreflective deposits situated between the RPE layer and the ellipsoid zone. SLO visualizes RPD as a reticular pattern of iso-reflective round lesions surrounded by a hyporeflective border (see Fig. 1).

[1] https://github.com/j-morano/multimodal-ssl-fpn.

2 Methods and Experimental Setup

The proposed approach, illustrated in Fig. 2, is as follows. Let $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^n$ and $\mathbf{y} \in \mathcal{Y} \subset \mathbb{R}^{n-1}$ be two images from modalities $\mathcal{X}$ and $\mathcal{Y}$, and $\mathbf{z} \in \mathcal{Z} \subset \mathbb{R}^{n-1}$, their corresponding target segmentation mask. Let images and masks from $\mathcal{Y}$ and $\mathcal{Z}$ have related anatomical features. We optimize a reconstruction model $\mathbf{y} = f_r(\mathbf{x}; \theta_r)$ and transfer its knowledge by initializing the weights of the segmentation model $\mathbf{z} = f_s(\mathbf{x}; \theta_s)$ with the optimized weights of the reconstruction model f_r. With this approach, modality $\mathcal{Y}$ serves as a free source of supervision, and images of this modality serve as soft segmentation targets. Thus, models can learn relevant patterns in a self-supervised way.

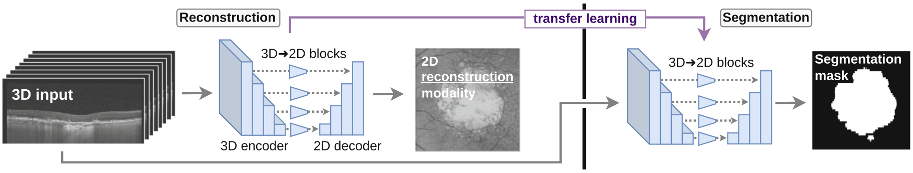

Fig. 2. Illustration of the proposed approach for 3D→2D segmentation. A novel 3D→2D model is trained for reconstructing image pairs of modalities with different dimensionality in a SSL setting, and then fine-tuned in the target segmentation task.

In this work, we propose a new CNN for the special case of 3D→2D segmentation, and we evaluate the proposed SSL approach for this case. In particular, we pre-train the new CNN to reconstruct SLO/FAF images from OCT (3D→2D), and then fine-tune it for GA/RPD segmentation. The advantage of reconstructing SLO over FAF is that several modern OCT devices allow to obtain co-registered OCT and SLO scans, providing the coordinates of each OCT slice within the SLO; thus, there is no need to use a separate registration method.

Network Architecture. The proposed network architecture (Fig. 3) is based on ReSensNet [20], and consists of a 3D encoder and a 2D decoder connected by novel 3D→2D feature projection blocks (FPBs). In the original work [20], training and inference are performed pixel-wise using fixed-size input patches. In contrast, we use full-size volumes of arbitrary resolution. To this end, we propose a novel type of FPB. In particular, all convolutions whose kernel size was equal to the expected feature size (calculated from the fixed size of the input patch) were replaced by $1 \times 1 \times 4$ convolutions. Then, to project 3D features at the output of each FPB to the 2D feature space, we add an adaptive average pooling of size 1 in the depth dimension at the end of each block. With this setting, feature selection and dimension reduction are performed at different scales, and the decoder processes only 2D features in the selected dimensions. This allows the model to learn the 3D structure of the data while being able to perform the segmentation in 2D. In addition, to overcome memory constraints and avoid overfitting, we reduce by half the number of kernels in each convolutional block.

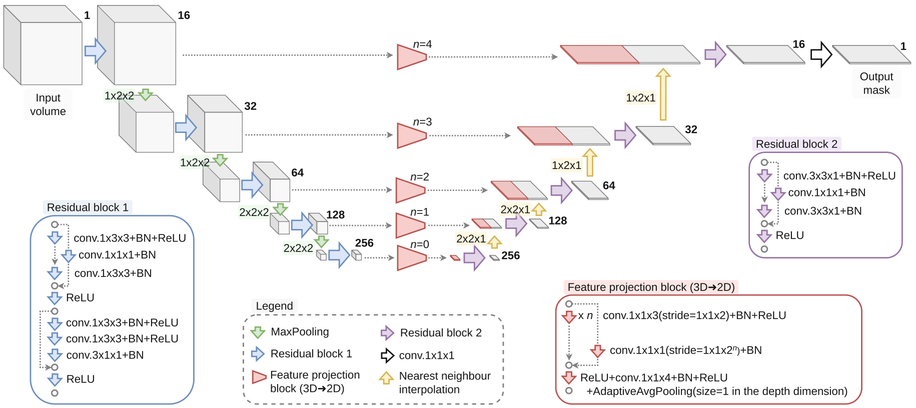

Fig. 3. Proposed 3D→2D CNN. Each residual encoder block has 8 3D convolutional layers, and each residual decoder block has 4 2D layers (number of feature maps also shown). The proposed feature projection block (FPB, in red) projects 3D features to the 2D feature space. FPBs have a variable number of $1 \times 1 \times 3$ convolutions followed by a $1 \times 1 \times 4$ convolution and a depth-wise adaptive average pooling of size 1. (Color figure online)

Table 1. Study data details. Dimensions: en-face height × en-face width × depth. FAF and SLO characteristics are after cropping to the same OCT en-face region projection.

Dataset	Scans	Patients (eyes)	Modality	Device	Area (mm)	Size (px)
GA-M	967	100 (184)	OCT	Spectralis	$6.68 \times 6.68 \times 1.92$	$49 \times 1024 \times 496$
			SLO	Spectralis	6.68×6.68	1024×1024
			FAF	Spectralis	6.68×6.68	1024×1024
GA-S	270	149 (166)	OCT	Spectralis	$6.02 \times 6.03 \times 1.92$	$49 \times 512 \times 496$
RPD-S	23	19 (23)	OCT	Spectralis	$5.73 \times 5.72 \times 1.92$	$97 \times 1024 \times 496$

Training Losses. As *reconstruction loss*, we use negative mean structural similarity index (NMSSIM) [23]. We empirically found that this loss performs equally or better than modern perceptual losses (e.g. LPIPS [25]) for our approach. NMSSIM loss can be defined as $\mathcal{L}_{\text{NMSSIM}}(\mathbf{x}, \mathbf{y}) = -\frac{1}{HW} \sum_{h,w} \text{SSIM}(\mathbf{x}_{hw}, \mathbf{y}_{hw})$, where $\mathbf{x}_{hw}$ and $\mathbf{y}_{hw}$ denote image patches of images $\mathbf{x}$ and $\mathbf{y}$ centered on the pixel with coordinates (h, w), $h \in H$ and $w \in W$, and $\text{SSIM}(\mathbf{x}_{hw}, \mathbf{y}_{hw})$ is the SSIM map for those patches, as described in [23]. As *segmentation loss*, we use the direct sum of Dice loss and Binary Cross-Entropy. These two losses are standard for binary segmentation tasks [8,9,11,12,16].

Datasets. Experiments were performed using three datasets (Table 1). *GA-M* samples come from a clinical study on GA progression. OCT and SLO images were automatically co-registered by the imaging device, while FAF images were registered with SLO using an in-house pipeline based on aligning retinal vessel segmentation. FAF and SLO images were cropped and resized to the same area and resolution as the OCT en-face projection. GA-M-S (35 samples) is a subset of GA-M with GA en-face masks annotated by a retinal expert on the OCT B-Scans. *GA-S* is composed of OCT volumes from another study with en-face GA annotations created by a retinal expert on the OCT B-scans. This dataset is divided patient-wise into two subsets: GA-S-2, containing volumes with annotations of two different experts, and GA-S-1, of only one. *RPD-S* is composed of OCT volumes with en-face RPD annotations created by retinal experts.

Training and Evaluation Details. OCT volumes were flattened along the Bruch's membrane, rescaled depth-wise to 128 voxels, and then Z-score normalized along the cross-sectional plane. To make FAF and GA masks more similar and thus facilitate fine-tuning, FAF images were inverted. In all cases, models were trained for 800 epochs using SGD with a learning rate of 0.1 and a momentum of 0.9. Batch size was set to 4 for reconstruction and 8 for segmentation.

All datasets were split patient-wise into training (60%), validation (10%) and test (30%). For reconstruction, models were trained on GA-M. For GA segmentation, they were trained/fine-tuned on GA-S-1 and evaluated on GA-S-1, GA-S-2 and GA-M-S. For RPD segmentation, RPD-S was used. To evaluate the performance under label scarcity, we train with 5%, 10%, 20% and 100% of the data in GA-S, and 20% and 100%, in RPD-S. More details about the hardware used and the carbon footprint of our method are included in the Supplement.

To reduce inference variability, we average the predictions of the top-5 checkpoints of the models in terms of Dice (validation). Segmentations are evaluated via Dice and absolute area difference (Area diff.) of predicted and manual masks.

3 Results and Discussion

Baseline Comparison. We compared our approach to current state-of-the-art methods (IPN [11], IPNv2 [12], Lachinov *et al.* [9], and ReSensNet [20]), showing that we greatly improve the state of the art in GA segmentation in scenarios with limited labeled data (Fig. 4). When using only 5% of the data (40 samples), the mean Dice score was 23% higher than the best state-of-the-art approach. Even without SSL, the proposed CNN improves the Dice score by 8%. This gain is even greater in terms of Area diff. The improvement is also visible in the predicted segmentation masks (Fig. 5). When using our approach, the number of false positives and negatives is highly reduced. On the other hand, the improvement for RPD segmentation is more modest (in this case, we only compared with the current state-of-the-art-method: Lachinov et al. [9]). This can be explained by the greater visibility of GA features compared to RPD features in FAF and SLO (see Figs. 1 and 5). This suggests that SSL benefits from images with similar pathomorphological manifestations.

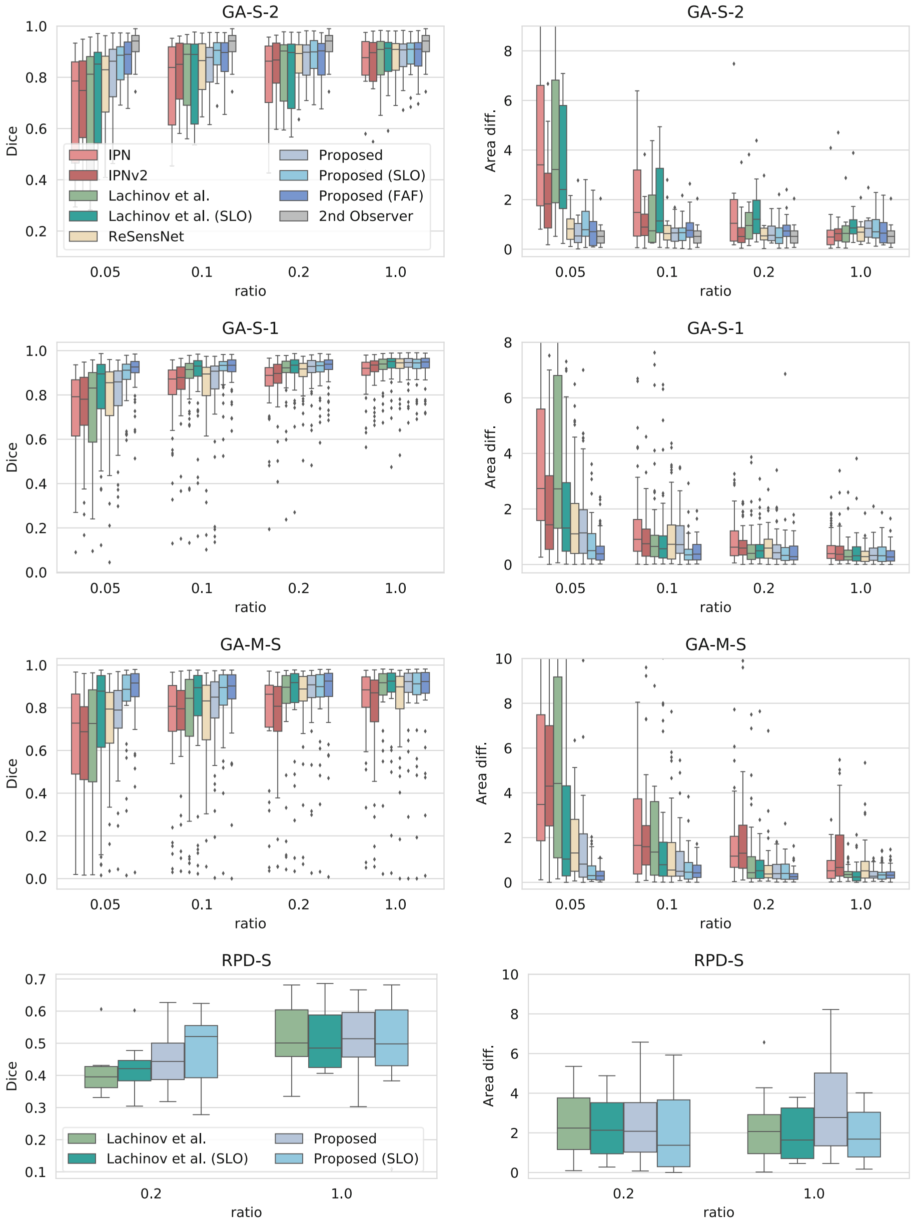

Fig. 4. Segmentation results of the models trained with different amounts of data. The title of each plot indicates the test dataset. If a model was pre-trained with SSL, the pre-training modality is shown in parentheses. A table with all means and standard deviations, as well as the results of a Wilcoxon signed rank test between our proposal and the others is included in the Supplement.

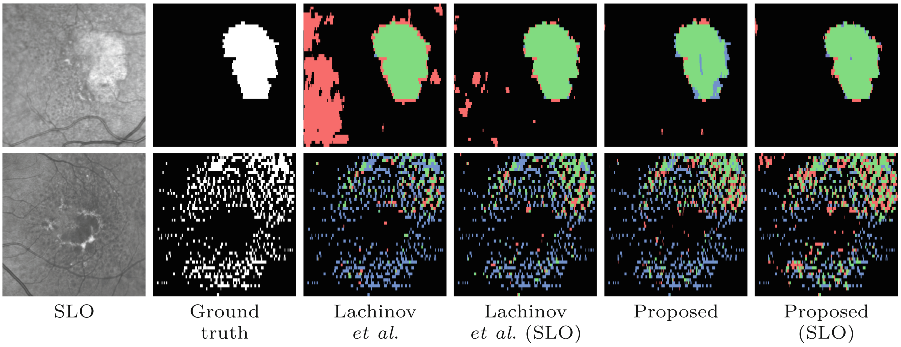

Fig. 5. Examples of GA (top) and RPD (bottom) segmentations from different models using the 5% and the 20% of the training data, respectively. True positives are depicted in green; true negatives, in black; false positives, in red; and false negatives, in blue. (Color figure online)

SSL Effect. To further assess the effect of the SSL, we also applied the strategy to the CNN by Lachinov et. al [9]. Figure 4 shows that SSL clearly improves the GA and RPD segmentation performance of both proposed and Lachinov *et al.* methods. These results are in line with the qualitative results in Fig. 5. This demonstrates that the SSL strategy is beneficial regardless of the architecture and the data. Notwithstanding, as discussed in the baseline comparison, the proposed SSL is more beneficial for GA segmentation than for RPD.

Reconstructed Modality Effect. We also conducted experiments to assess the effect of the reconstructed modality (SLO and FAF). Figure 4 shows that using FAF for SSL usually leads to better segmentation performance than using SLO. However, in multiple cases, the differences were not statistically significant. This is important because, unlike FAF, SLO does not require an external registration method.

4 Conclusions

Labeled data scarcity is one of the main limiting factors for the application of deep learning in medical imaging. In this work, we have proposed a new model and SSL strategy for label-efficient 3D→2D segmentation. The proposed approach was validated in two tasks with clinical relevance: the en-face segmentation of GA and RPD in OCT. The results demonstrate that: (1) the proposed CNN architecture clearly outperforms the state of the art when there is limited annotated data, (2) regardless of the architecture and the modality to be reconstructed, the proposed SSL strategy improves the performance of the models on the target tasks in those cases; (3) despite the greater diagnostic utility of FAF over SLO, SSL with FAF does not always result in a significant gain in model performance, with the advantage of the latter not requiring a supplementary registration method. On the other hand, although the proposed approach shows promising results in the en-face segmentation of RPD, further evaluation is needed.

Based on our findings, we believe that the proposed approach has the potential to be used in other common 3D→2D tasks, such as the prediction of retinal sensitivity in OCT, the segmentation of different structures in OCT-A, or the segmentation of intravascular ultrasound (IVUS). In addition, we also believe that the proposed SSL strategy could be easily extended to other imaging domains, such as magnetic resonance, where multi-modal data is widely used.

Acknowledgements. This work was supported in part by the Christian Doppler Research Association, Austrian Federal Ministry for Digital and Economic Affairs, the National Foundation for Research, Technology and Development, and Heidelberg Engineering.

References

1. Kalapos, A., Gyires-Tóth, B.: Self-supervised pretraining for 2D medical image segmentation. In: Karlinsky, L., Michaeli, T., Nishino, K. (eds.) ECCV 2022. LNCS, vol. 13807, pp. 472–484. Springer, Heidelberg (2022). https://doi.org/10.1007/978-3-031-25082-8_31
2. Brempong, E.A., Kornblith, S., Chen, T., Parmar, N., Minderer, M., Norouzi, M.: Denoising pretraining for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4175–4186 (2022)
3. Bui, P.T.A., et al.: Fundus autofluorescence and optical coherence tomography biomarkers associated with the progression of geographic atrophy secondary to age-related macular degeneration. Eye **36**(10), 2013–2019 (2021). https://doi.org/10.1038/s41433-021-01747-z
4. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., et al.: Bootstrap your own latent: a new approach to self-supervised learning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS 2020. Curran Associates Inc. (2020)
5. He, K., Girshick, R., Dollar, P.: Rethinking ImageNet pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
6. Hervella, Á.S., Rouco, J., Novo, J., Ortega, M.: Retinal image understanding emerges from self-supervised multimodal reconstruction. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 321–328. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_37
7. Hervella, A.S., Rouco, J., Novo, J., Ortega, M.: Learning the retinal anatomy from scarce annotated data using self-supervised multimodal reconstruction. Appl. Soft Comput. **91**, 106210 (2020). https://doi.org/10.1016/j.asoc.2020.106210
8. Kavur, A.E., Gezer, N.S., Barış, M., Aslan, S., Conze, P.H., Groza, V., et al.: CHAOS challenge - combined (CT-MR) healthy abdominal organ segmentation. Med. Image Anal. **69**, 101950 (2021). https://doi.org/10.1016/j.media.2020.101950

9. Lachinov, D., Seeböck, P., Mai, J., Goldbach, F., Schmidt-Erfurth, U., Bogunovic, H.: Projective skip-connections for segmentation along a subset of dimensions in retinal OCT. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 431–441. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_41
10. Lee, K., Zung, J., Li, P., Jain, V., Seung, H.S.: Superhuman accuracy on the SNEMI3D connectomics challenge (2017). https://doi.org/10.48550/ARXIV.1706.00120
11. Li, M., et al.: Image projection network: 3D to 2D image segmentation in OCTA images. IEEE Trans. Med. Imaging **39**(11), 3343–3354 (2020). https://doi.org/10.1109/TMI.2020.2992244
12. Li, M., et al.: OCTA-500: a retinal dataset for optical coherence tomography angiography study (2022)
13. Liefers, B., González-Gonzalo, C., Klaver, C., van Ginneken, B., Sánchez, C.I.: Dense segmentation in selected dimensions: application to retinal optical coherence tomography. In: Cardoso, M.J., et al (eds.) Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning. Proceedings of Machine Learning Research, vol. 102, pp. 337–346. PMLR (2019)
14. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging **34**(10), 1993–2024 (2015). https://doi.org/10.1109/TMI.2014.2377694
15. Morano, J., Álvaro S. Hervella, Barreira, N., Novo, J., Rouco, J.: Multimodal transfer learning-based approaches for retinal vascular segmentation. In: Giacomo, G.D., et al. (eds.) Proceedings of the 24th European Conference on Artificial Intelligence (ECAI 2020), pp. 1866–1873 (2020). https://doi.org/10.3233/FAIA200303
16. Orlando, J.I., Fu, H., Barbosa Breda, J., van Keer, K., Bathula, D.R., Diaz-Pinto, A., et al.: REFUGE challenge: a unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Med. Image Anal. **59**, 101570 (2020). https://doi.org/10.1016/j.media.2019.101570
17. Raghu, M., Zhang, C., Kleinberg, J., Bengio, S.: Transfusion: understanding transfer learning for medical imaging. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc. (2019)
18. Ross, T., et al.: Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. Int. J. Comput. Assist. Radiol. Surg. **13**(6), 925–933 (2018). https://doi.org/10.1007/s11548-018-1772-0
19. Schmitz-Valckenberg, S., et al.: Natural history of geographic atrophy progression secondary to age-related macular degeneration (geographic atrophy progression study). Ophthalmology **123**(2), 361–368 (2016). https://doi.org/10.1016/j.ophtha.2015.09.036
20. Seeböck, P., et al.: Linking function and structure with ReSensNet: predicting retinal sensitivity from OCT using deep learning. Ophthalmol. Retina **6**(6), 501–511 (2022). https://doi.org/10.1016/j.oret.2022.01.021
21. Sun, S., Sonka, M., Beichel, R.R.: Graph-based IVUS segmentation with efficient computer-aided refinement. IEEE Trans. Med. Imaging **32**(8), 1536–1549 (2013). https://doi.org/10.1109/TMI.2013.2260763
22. Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J.N., Wu, Z., Ding, X.: Embracing imperfect datasets: a review of deep learning solutions for medical image segmentation. Med. Image Anal. **63**, 101693 (2020). https://doi.org/10.1016/j.media.2020.101693

23. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
24. Wei, W., et al.: Two potentially distinct pathways to geographic atrophy in age-related macular degeneration characterized by quantitative fundus autofluorescence. Eye (2023). https://doi.org/10.1038/s41433-022-02332-8
25. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 586–595 (2018). https://doi.org/10.1109/CVPR.2018.00068