Supplementary Material of Harvard Glaucoma Detection and Progression: A Multimodal Multitask Dataset and Generalization-Reinforced Semi-Supervised Learning

Table 1: Performance of different *supervised methods* on the cross-sectional data with single modality RNFLT for the *glaucoma detection* task on the released **1,000** glaucoma detection data.

Model	**Acc↑**	**F1↑**	**AUC↑**
VGG [3]	0.80	0.79	0.86
ResNet [4]	0.84	0.83	0.87
ResNext [5]	0.82	0.81	0.89
WideResNet [6]	0.83	0.84	0.89
EfficientNet[7]	0.85	0.85	0.90
ConvNext [8]	0.80	0.79	0.86
ViT [9]	0.65	0.67	0.75
Swin [10]	0.74	0.73	0.78

1. Implementation Detail for Supervised Benchmarks

For the optimization, we use AdamW optimizer [1] and train all the supervised models with 20 epochs throughout all the experiments. We use learning rate 2e-5 and weight decay 1e-5 with a batch size of 12 for all the supervised classification models for all methods in supervised progression forecasting and glaucoma detection benchmarks. All supervised classification models are trained using BCE loss. For ViT, we used their ViT-B-16 architecture. For EfficientNet, we use their EfficientNetV2-S architecture. For Swin transformer, we use their Swin-base architecture. For ResNet, we use their ResNet50 architecture. For VGG, we use their VGG-11 architecture. For ResNeXt, we use their ResNeXt-101 64×4d architecture. For WiderResNet, we use their Wide ResNet-50-2 architecture. For ConvNeXt, we use their ConvNeXt Tiny architecture. We initialize all models with pre-trained imagenet weights. All code is written in PyTorch [2] and we use one RTX A6000 GPU for all experiments.

2. Supervised Benchmarks on Released Data

In Table1, we show the supervised classification results for the **glaucoma detection** task with multiple SOTA supervised CNN and transformer baseline methods, including VGG [3], ResNet [4], ResNext [5], WideResNet [6], EfficientNet [7], ConvNext [8], ViT [9], and Swin Transformers [10]. This **cross-sectional** benchmark is conducted on our future cross-sectional data release with 1,000 patients upon acceptance, of which 800 patients are used for training and the remaining 200 are used for testing. To the best of our knowledge, this is the largest supervised glaucoma detection benchmark with 3D OCT imaging data (i.e., RNFLT). Such large-scale public 3D OCT dataset will encourage researchers to build clinically effective (3D OCT source data) and efficient (post-processed 2D RNFLT map from the 3D data) glaucoma CAD systems. As shown in the table, transformed-based architectures tend to obtain worse performance than CNN-based architectures, and we articulate this due to that transformed-based architectures are often data-hungry and require a relatively larger amount of training data. EfficientNet is the best-performing method, followed by WideResNet and ResNext.

2.1. Data Density Distribution

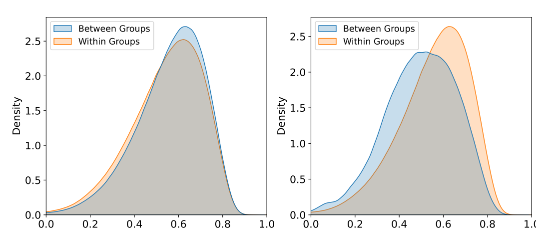

Figure 1: The pointwise similarity between RNFLT maps within the same label groups versus the pointwise similarity between RNFLT maps between different label groups.

As shown in Fig. 1, the density distribution of correlations between RNFL thickness (RNFLT) maps within groups of glaucoma and non-glaucoma is largely overlapped with the one between RNFLT maps between glaucoma and non-glaucoma groups. The same is observed for progression versus non-progression.

References

[1] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 1

[2] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary 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. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 1

[3] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[5] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 1

[6] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 1

[7] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 1

[8] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 1

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[10] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 1