Supplementary Material

1 EXPERIMENTAL RESULTS ON CLEAN APTOS 2019

Proper orthogonality constraints can help fully utilize the model capacity. We also have conducted ablation experiments for each part of the framework based on the ViT3 model.

The experimental results show that adding soft orthogonal constraints to all layers of DNN can help improve the model performance. The experimental results on the clean dataset are summarized in Tab. 1. It can be seen that the proposed TAOTF-based models have better performances than other methods. Additionally, as it is an imbalanced dataset, we also conduct the ablation experiments on the clean dataset and evaluate the performances with Recall and Precision, which can be seen in Tab. 1.

	Accuracy	**Precision**	**Recall**
ResNet	91.17%	91.06%	90.93%
ResNet+hard constraints	90.75%	90.62%	90.53%
ResNet+SRIP	91.17%	91.10%	91.17%
ResNet+OCNN	91.08%	91.26%	91.08%
ResNet+WaveCNet	92.18%	92.34%	92.18%
TAOTF-ResNet (Ours)	**92.53%**	**92.46%**	**92.53%**
ViT3	90.99%	90.99%	90.83%
ViT3+orth-initialization	89.92%	89.85%	89.49%
ViT3+hard constraints	87.62%	87.58%	87.29%
ViT3+SRIP	91.07%	91.09%	91.02%
ViT3+stage2 (only self-attention layers)	92.60%	92.78%	92.21%
ViT3+stage2 (only transformer blocks)	95.74%	95.78%	95.71%
ViT3+stage2 (only patch embedding)	95.11%	95.16%	95.00%
TAOTF-ViT3 (Ours)	**95.87%**	**95.81%**	**95.80%**

Table 1: Results on clean Kaggle APTOS 2019 test sets (Ablation experiments). Our framework imposes models with proper orthogonality constraints, which can improve task performance. For compared methods, we set the same hyperparameters for a fair comparison.

2 EXPERIMENTAL RESULTS ON GLAUCOMA DETECTION (RECALL)

Due to the presence of imbalanced datasets, evaluating the performance of machine learning models solely based on accuracy may not be an appropriate approach. In such datasets, the number of instances in one or more classes may be significantly larger or smaller than the others. As a result, a model may achieve high accuracy by simply predicting the majority class, while ignoring the minority classes.

In the field of image classification, recall is a commonly used evaluation metric that measures the ability of a model to correctly identify all positive instances in a dataset. Recall is particularly useful when dealing with imbalanced datasets, as it provides an indication of how well a model is able to detect the minority classes in the dataset. A high recall value indicates that the model is able to correctly identify a large proportion of the positive instances in the dataset, regardless of the size of the minority class.

To more appropriately evaluate the performance of TAOTF, we also evaluate its performance using recall. The results of this evaluation can be seen in Tab. 2. By evaluating the model using both accuracy and recall, we can gain a better understanding of its overall performance, particularly in datasets that are imbalanced.

Experiment on CIFAR-100													
	Clean	Noise			Blur			Weather			Digital		
Methods	Clean	Gaussian.	ISO.	Multiplicative.	Gauss.	Median	Motion	Optical	Rotate	RGB	Bright	Frog	Saturation
WideResnet	68.87%	35.87%	21.75%	27.77%	03.86%	13.25%	19.25%	26.50%	23.34%	51.93%	52.19%	54.07%	47.59%
WideResnet+SRIP	70.97%	45.06%	33.02%	40.76%	07.40%	19.29%	29.31%	37.53%	41.24%	59.06%	58.92%	60.95%	55.19%
WideResnet+hard constraints	71.04%	53.40%	38.24%	46.72%	17.94%	27.70%	41.60%	52.18%	42.85%	64.54%	64.06%	65.91%	60.20%
TAOTF-WideResnet (Ours)	**71.09%**	**61.06%**	**45.66%**	**56.02%**	**20.24%**	**33.89%**	**47.15%**	**57.56%**	**59.04%**	**69.52%**	**69.21%**	**71.01%**	**65.35%**

Table 2: Experimental Results on Glaucoma Detection (Recall).

3 EXPERIMENTAL RESULTS ON BRAIN MRI SEGMENTATION

To ensure statistical significance, experiments should be conducted multiple times and the results should be statistically analyzed. In this study, all experimental results exceeded the average of twenty test runs, demonstrating their reliability and consistency.

However, the accuracy in Table 4 is marginal. To provide a more detailed and comprehensive analysis of the experimental results, we present additional experimental results in Tab. 3.

	BCE (↓)	**ACC (↑)**	**F1-Score (↑)**
UNet	0.00713 (0.00709-0.00726)	0.996 (0.9952-0.9965)	0.837 (0.829-0.841)
UNet+SRIP	0.00949 (0.00873-0.00971)	0.996 (0.9960-0.9971)	0.811 (0.808-0.817)
TAOTF-UNet (Ours)	**0.00697** **(0.00692-0.00701)**	**0.997** **(0.9970-0.9976)**	**0.843** **(0.841-0.845)**

Table 3: Results on Brain MRI segmentation test sets. The proposed TAOTF can help improve segmentation performance.