ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230310

TAOTF: A Two-Stage Approximately Orthogonal Training Framework in Deep Neural Networks

Taoyong Cui[1], Jianze Li[2], Yuhan Dong[1] and Li Liu[3] ✉

[1]Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[2]Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, Guangdong, China
[3]The Hong Kong University of Science and Technology (Guangzhou), China
✉Corresponding author. avrillliu@hkust-gz.edu.cn

Abstract. The orthogonality constraints, including the hard and soft ones, have been used to normalize the weight matrices of Deep Neural Network (DNN) models, especially the Convolutional Neural Network (CNN) and Vision Transformer (ViT), to reduce model parameter redundancy and improve training stability. However, the robustness to noisy data of these models with constraints is not always satisfactory. In this work, we propose a novel two-stage approximately orthogonal training framework (TAOTF) to find a trade-off between the orthogonal solution space and the main task solution space to solve this problem in noisy data scenarios. In the first stage, we propose a novel algorithm called polar decomposition-based orthogonal initialization (PDOI) to find a good initialization for the orthogonal optimization. In the second stage, unlike other existing methods, we apply soft orthogonal constraints for all layers of DNN model. We evaluate the proposed model-agnostic framework both on the natural image and medical image datasets, which show that our method achieves stable and superior performances to existing methods. Supplementary materials can be found in https://github.com/nonameinformation/anonymous/tree/main.

1 Introduction

In the past decades, *Deep Neural Network* (DNN) models, especially the *Convolutional Neural Network* (CNN) and *Vision Transformer* (ViT), have developed rapidly in the computer vision field [19, 30, 31, 25]. Although these models can automatically learn the hidden deep features from images, there still exist several problems with them. For example, the parameterization or model capacity utilization is insufficient, gradient explosion or disappearance, and there exists significant redundancy among different feature channels [26]. In view of this, orthogonality constraints, including the hard and soft ones, were recently used in the field of deep learning to improve model performance. When the filters are learned to be as orthogonal as possible, they become irrelevant and reduce the redundancy of learning features [26]. Then the model capacity is made full use of, and the ability of feature expression is improved as well. For example, a hard orthogonality constraint was imposed in CNN [10], and retraction-based Riemannian optimization algorithms were used to solve it. A soft orthogonality constraint was imposed in CNN [26] with an orthogonal penalty loss.

However, in these works, inappropriate orthogonal constraints are often imposed, ignoring a more important advantage of orthogonal constraints: robustness to noise samples (*e.g.*, noise, blur, exposure, and so on). Applying appropriate orthogonal constraints can make each layer of the model closer to a 1-Lipschitz function. Given a small perturbation to input Δx, the change of output Δy is bounded to be low. Therefore, the model enjoys robustness under noisy data. For example, if we only use the hard orthogonality constraint, one issue is that the solution set of the primary optimization objective does not necessarily intersect with the hard orthogonality constraints [6]. In other words, if the weight matrices of these models are too close to orthogonal matrices, the performance may be worse. Meanwhile, the computing cost of the hard constraints is always expensive. On the other hand, if we only use the soft orthogonality constraint, it is ineffective to make the weight matrix orthogonal enough, and thus reduces layers' 1-Lipschitz property. More detailed explanations about these issues will be presented in Sec. 3.1.

In this work, to solve the above issues, we propose a *two-stage approximately orthogonal training framework* (TAOTF) to find the trade-off between the Stiefel manifold and the main task solution set. More specifically, in the first stage, we propose a novel algorithm called *polar decomposition-based orthogonal initialization* (PDOI) to find a starting point in the Stiefel manifold. This process is somewhat similar to the hard orthogonal constraints but with a much smaller computational cost. Then, in the second stage, we implement soft orthogonal regulation with an orthogonal penalty loss (*e.g.*, *spectral restricted isometry property* (SRIP) [2]) on all layers of DNN, and use a common European optimizer (*e.g.*, Adam) to find an optimal point. We train the CNN and ViT models using the proposed TAOTF framework and then evaluate these TAOTF-based models on both natural and medical images.

To evaluate the robustness of our framework compared to other methods, we simulate possible data challenges with datasets and conduct comparative experiments, which demonstrate the superiority of our framework. Except for these two models, this novel framework can be also used together with other DNN models, *e.g.*, the *Recurrent Neural Network* (RNN), to further improve the robustness performances in real scene datasets.

In summary, three contributions can be summarized as follows:

(1) We propose a novel model-agnostic framework called TAOTF by combining the advantages of soft and hard orthogonality constraints to improve the robustness performances in DNNs, especially the CNN and ViTs.
(2) In the first stage of TAOTF, to find a suitable starting point for or-

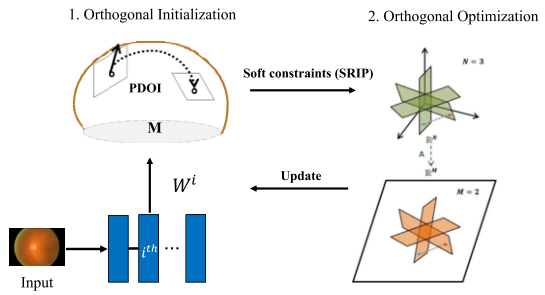

Figure 1: Our proposed two-stage orthogonal training framework (TAOTF). In the first stage, we use PDOI to find a suitable starting point on Stiefel Manifold by iterating with low computing costs. In the second stage, we impose soft orthogonal constraints on all layers of DNN to find a trade-off.

thogonal optimization, we propose a novel algorithm called PDOI to search near the initial point and update parameter matrices. To our knowledge, this is the first time that a projection-based Riemannian optimization algorithm is used in the training of DNNs.

(3) We also conduct extensive experiments in many image datasets, including natural and medical ones. The experimental results show that TAOTF-based models have better robustness performances to noisy perturbations than existing methods.

2 Related Works

In this section, we mainly review some related works in the literature about the applications of hard and soft orthogonality constraints in the DNN models, especially the CNN and Transformers.

2.1 Hard orthogonality constraints

To our knowledge, the first work using hard orthogonality constraints in CNN was [10], where the Stiefel layer was introduced, and Riemannian optimization techniques on matrix manifolds were used in AlexNet and VGG. Then, a new backpropagation with a variant of *stochastic gradient descent* (SGD) on Stiefel manifolds [13] was exploited to update the structured connection weights. In [12], the authors generalized such square orthogonal matrices to rectangular ones, and formulated this problem in *Feed-forward Neural Networks* (FNNs) as an optimization problem over multiple dependent Stiefel manifolds. Recently, in [14], an alternative approach was proposed based on a parameterization stemming from Lie group theory, and the constrained optimization problem was transformed into an unconstrained one over a Euclidean space.

2.2 Soft orthogonality constraints

The soft orthogonality constraints were mainly solved using a penalty loss function calculating the discrepancy between the identity matrix $\boldsymbol{I}$ and the product of weight matrix $\boldsymbol{W}$ and its transpose $\boldsymbol{W}^{\boldsymbol{T}}$, *e.g.*, $\|\boldsymbol{W}\boldsymbol{W}^{\boldsymbol{T}} - \boldsymbol{I}\|$. To our knowledge, the first work using this soft orthogonality constraint in training DNNs was [21], where it was used in RNNs to help avoid gradient vanishing/ explosion, and the first work using the soft orthogonality constraint in CNNs was [27], which helped to stabilize the layer-wise distribution of activations. In [2], to enforce the orthogonality regularizations, the authors used a novel regularization form for orthogonality in CNNs, named *Spectral Restricted Isometry Property* (SRIP). In [26], a new *orthogonality based CNN* (OCNN) was proposed, and good results on multiple natural datasets were achieved approving their robustness under attack. In [7], class vectors were applied to improve the ability of the model to resist the label noise of datasets for cancer diagnosis.

In transformers, orthogonal weights can also improve numerical stability during training and upper-bound the Lipschitz constant of linear transformations. In [29], they first applied the basic orthogonality constraint on transformers and achieved good results in several NLP tasks such as neural machine translation and sequence-to-sequence dialogue generation. In [6], they developed an orthogonal Vision Transformer (O-ViT), which also used methods like [14] to impose orthogonality constraints on self-attention layers.

3 Methods

In this section, we will detailedly introduce the new training framework TAOTF (Fig. 1). In Sec. 3.1, we will explain the reason why we propose a new two-stage orthogonal training framework. In Sec. 3.2, the proposed algorithm PDOI to find a good optimization starting point in the first stage will be introduced. In Sec. 3.3, we will introduce the soft constraints we use in the second stage.

3.1 Why we need a two-stage framework?

In a DNN model, it is well known that, if $\boldsymbol{X}$ is the weight matrix of the i-th layer, then the training of $\boldsymbol{X}$ is to solve the following optimization problem:

$$\min_{\boldsymbol{X}\in\mathbb{R}^{n\times p}} g(\boldsymbol{X}) \tag{1}$$

where g is the loss function. Let

$$\mathbf{St}(p,n) \stackrel{\text{def}}{=} \{\boldsymbol{X}\in\mathbb{R}^{n\times p} : \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} = \boldsymbol{I}_p\}$$

be the *Stiefel manifold*, where $1 \le p \le n$, and $\boldsymbol{I}_p$ denotes the identity matrix of size p. As explained in Sec. 1, to reduce the redundancy of learning features, we would like to impose orthogonality on the weight matrix $\boldsymbol{X}$.

One approach is to use a *hard orthogonality constraint*, and then problem (1) becomes a Riemannian optimization problem [1] on the Stiefel manifold [15], *i.e.*,

$$\min_{\boldsymbol{X}\in\mathbf{St}(p,n)} g(\boldsymbol{X}). \tag{2}$$

In the iterations of the algorithm to solve problem [1], the weight matrix $\boldsymbol{X}$ will always stay in $\mathbf{St}(p,n)$, and thus can be kept columnly orthogonal. The other approach is to use a *soft orthogonality constraint*, and then problem (1) becomes

$$\min_{\boldsymbol{X}\in\mathbb{R}^{n\times p}} g(\boldsymbol{X}) + \lambda r(\boldsymbol{X}), \tag{3}$$

where $\lambda > 0$ and $r(\boldsymbol{X})$ is a regularization term to enforce the orthogonality of $\boldsymbol{X}$.

As introduced in Sec. 2, the above hard and soft orthogonality constraints were both used to improve the robustness performances of DNN or ViT models. However, as explained in Fig. 2, the solution matrices of problem (1) maybe not be in $\mathbf{St}(p,n)$. Therefore, if we only use the hard orthogonality constraint to solve problem (2), the solution may be too restrictive. In other words, the solution set of the primary optimization objective does not necessarily intersect with

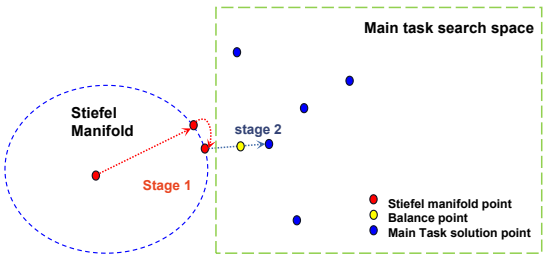

Figure 2: Stiefel manifold and the solution set. The solution set of the primary optimization objective does not necessarily intersect with the Stiefel manifold, our method can find a balance point between them.

the hard orthogonality constraints [6], and thus the performance of the trained DNN models may degrade. On the other hand, if we only use the soft orthogonality constraint to solve problem (3), the solution may be far away from $\mathbf{St}(p, n)$, and thus the DNN models may still suffer from the parameter redundancy, and it is not robust enough as well.

Enforcing proper orthogonality constraints can generate a more uniform spectrum of $\boldsymbol{W}$ [26], which makes network layers (like the convolution layer f) closer to a 1-Lipschitz function like

$$\|f(x_1) - f(x_2)\| \leq \|x_1 - x_2\|. \tag{4}$$

And in the analysis of numerical stability, enforcing orthogonality constraints can upper-bound the Lipschitz constant of linear transformations [29]. The Lipschitz constant is a measure that estimates the rate of change (variation) of representations. Given a slight perturbation to the input Δx, the variation of the output Δy is bounded low, producing a robust and less sensitive representation of data perturbations. Therefore, the model enjoys robustness to noisy data.

Therefore, in this work, we propose a novel TAOTF framework, which includes two stages at each iteration to find the trade-off between the search space of the main task and the orthogonality constraint. From a perspective of optimization theory, it can be understood that we first solve problem (2) to calculate a suitable starting point, and then solve problem (3). It will be seen in Sec. 4 that, although the TAOTF framework includes two stages, it still has competitive convergence speed. One reason is the use of a projection-based PDOI algorithm in the first stage, which does not need to calculate the retraction map. The other reason is that we control the iteration numbers at the first stage.

3.2 First Stage: Orthogonal Initialization

To solve the problem (2), the retraction-based optimization algorithm [1] was proposed in recent years. However, as the retraction-based algorithms are generally expensive, in this work, inspired by low-rank orthogonal approximation of tensors [4], we propose a novel algorithm PDOI in Algorithm 1, to find a local optimum as the starting point for orthogonal optimization of the second stage. Although the global convergence of the X_k is not determined, under this mild condition, the input point can be converged (locally) to an extreme point by the PDOI algorithm nearby, which can be used as the starting point for the next stage. The initial point found in this way, on the one hand, satisfies the constraint of the Stifel manifold, and on the other hand, finds a point closer to the main task search space on the Stifel manifold with low computing costs. We use the SVD to compute the polar decompositions (i.e., projection). Compared with other retractions (such as exponential mapping), polar decomposition is a simpler way and not expensive. Furthermore, as it is an initialization method, we only need to iterate a few times in the first stage.

Algorithm 1: PDOI algorithm

Input: a starting point $\boldsymbol{X}_0$, a positive constant $\gamma > 0$.
Output: $\boldsymbol{X}_k$, $k \geq 1$.
1: **for** k=1,2,..., until a stopping criterion is satisfied **do**
2: Compute $\nabla g(\boldsymbol{X}_{k-1})$.
3: Compute the SVD decomposition ($\boldsymbol{U\Sigma V^T}$) of

$$\nabla g(\boldsymbol{X}_{k-1}) + \gamma \boldsymbol{X}_{k-1}. \tag{5}$$

4: Update $\boldsymbol{X}_k$ to be the product of two orthogonal matrices $\boldsymbol{UV^T}$.
5: **end for**

The PDOI algorithm employs an alternating procedure (iterating through X_0, X_1,...,X_k,...,X_N), where in each step all but one (X_n) parameters are fixed [4]. In general, algorithms of this type, including alternating least squares, are not guaranteed global convergence, but the iterations can search for points closer to the main task on the Stiefel manifold. Moreover, of the generated parameter sequence, every converging subsequence converges to a stationary point of the objective function, which can be a suitable starting point from the perspective of optimization theory.

As proved in [11], the initial weights from the orthogonal group not only speeds up convergence relative to the standard Gaussian initialization with iid weights but close to isometry during training to enable efficient convergence and the 1-Lipschitz property. The algorithm PDOI guarantees that the starting point is on the Stiefel manifold (initial weights in the orthogonal group), and through multiple iterations, it can find a better starting point close to the input point that is suitable for both the orthogonal constraints and the main tasks. Extensive experiments have been conducted to prove this view.

3.3 Second Stage: Orthogonal Optimization

Recall that the *Restricted Isometry Property* (RIP) condition of a matrix $\boldsymbol{W} \in \mathbb{R}^{n\times n}$ means that, for all vectors $\boldsymbol{x} \in \mathbb{R}^n$ that are k-sparse, there exists a small $\delta_{\boldsymbol{W}} \in (0, 1)$ such that

$$(1 - \delta_{\boldsymbol{W}}) \|\boldsymbol{x}\|_2^2 \leq \|\boldsymbol{W}\boldsymbol{x}\|_2^2 \leq (1 + \delta_{\boldsymbol{W}}) \|\boldsymbol{x}\|_2^2. \tag{6}$$

The positive constant $\delta_{\boldsymbol{W}}$ in equation (6) is called the *constrained isometric constant*. If δ_W is very small, it can be interpreted as those k columns are approximately orthogonal. If the equation (6) is satisfied with $\delta_{\boldsymbol{W}} = 0$ for all k-sparse vectors $\boldsymbol{x} \in \mathbb{R}^n$, then the matrix $\boldsymbol{W} \in \mathbb{R}^{n\times n}$ satisfies the *isometric characteristics of k-order constraints*. The RIP can be used to measure the similarity between the subset composed of k columns in a matrix and an orthogonal matrix.

The extreme case with $k = n$ was also considered in [2], where the RIP condition will force the whole matrix $\boldsymbol{W}$ to be very close to an orthogonal one, *i.e.*,

$$\left| \frac{\|\boldsymbol{W}\boldsymbol{x}\|^2}{\|\boldsymbol{x}\|^2} - 1 \right| \leq \delta_{\boldsymbol{W}}, \forall \boldsymbol{x} \in \mathbb{R}^n. \tag{7}$$

In this case, the RIP condition (7) is termed as the *Spectral Restricted Isometry Property* (SRIP) regularization.

Existing methods [26, 29, 2, 6, 10, 8] always only impose orthogonality constraints on the deep convolutional layers or the self-attention layer of Transformers. However, if we only impose orthogonality constraints on only some layers of the network, the different levels of orthogonality will destroy the 1-Lipschitz property of the global model and damage model robustness against small perturbations. And we proved this view through extensive experiments in Sec. 4. Therefore, we impose soft constraints like SRIP regulation (7) on all layers of the model to solve the orthogonal optimization in the second stage.

This process is to separate the starting point found in the first stage from the Stiefel manifold, and further explore the main task solution space but is still limited by the Stiefel manifold. Through such a process, the trade-off of the main task and the orthogonal constraints can be well found, and the final loss function in (3) is

$$\min_{\boldsymbol{X}\in\mathbb{R}^{n\times p}} g(\boldsymbol{X}) = g_M(\boldsymbol{X}) + \lambda g_{SRIP}(\boldsymbol{X}), \tag{8}$$

where $g_{SRIP}(\boldsymbol{X})$ is the spectral norm of $\boldsymbol{W}\boldsymbol{W}^{\boldsymbol{T}} - \boldsymbol{I}$, that is $\sup_{\boldsymbol{x}\in\mathbb{R}^n,\boldsymbol{x}\neq 0}\left|\frac{\|\boldsymbol{W}\boldsymbol{x}\|^2}{\|\boldsymbol{x}\|^2} - 1\right|$, and $g_M(\boldsymbol{X})$ is main task loss.

4 Experiments

In this section, to evaluate the efficiency of the proposed TAOTF framework, we conduct several experiments of the TAOTF-based DNN models on various datasets, including natural and medical ones. We implement these models on top of the deep learning framework PyTorch. Unless otherwise stated, the experimental results are measured in Top-1 Accuracy. We train the models with the clean dataset and test them with simulated noisy datasets. For simulating noise, we utilize the Albumentations tool [3], with the simulated noise intensity as its default initial value. All experimental results in this work exceed the average of twenty test results. More detailed experimental results can be seen in Supplementary Material.

4.1 Experiments on Kaggle APTOS 2019

4.1.1 Dataset

We first use the public dataset *Kaggle APTOS 2019*, which was collected by the Aravind Eye Hospital in India's rural areas, to evaluate the proposed TAOTF-CNN and TAOTF-ViT models.

This dataset contains 3662 retinal images, and the labels were provided by the clinicians who rated the development of Diabetic retinopathy (DR) in each image by a scale of "0, 1, 2, 3, 4", meaning "*no DR*", "*mild*", "*moderate*", "*severe*" and "*proliferative DR*", respectively. Note that this dataset doesn't have equal distributions among the different classes. For example, it has far more normal data with the label "0" than other classes. We randomly shuffle the entire dataset into three subgroups, *i.e.*, training (70%), validation (10%), and testing (20%).

4.1.2 Image Preprocessing

As different fundus images have different length-width ratios, and the width of different black edges around the eyeball is also different, we can not straightly resize the images based on their sizes. Therefore, in this experiment, we resize the fundus images (224×224) based on the eyeball radius and then use the feature enhancement method. In this process, the difference between the original image and the Gaussian blurred one (equivalent to the background) is used to enhance the feature.

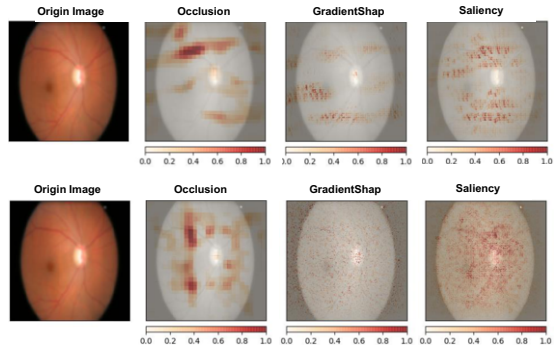

Figure 3: The Glaucoma Detection Classification visualization (blur image) by Towhee, including Occlusion, GradientShap, and Saliency. And the first row is the classification visualization of the baseline model ResNet, and the second row is the classification visualization of TAOTF-based ResNet.

4.1.3 Models and Settings

We choose the ResNet18, ViT3 (3 transformer blocks), and ViT6 (6 transformer blocks) models to test the robustness performances of TAOTF-based models on the Diabetic Retinopathy classification task. For training these models, the total epoch of the training is 200. We start the learning rate with $lr = 3\times10^{-5}$, with weight decay 1e-4. The weight λ of the regularization loss is 10^{-3}, the model is trained using Adam, and the batch size is 8. In the above training process, we use Cross Entropy (CE) Loss for the criterion.

4.1.4 Experimental Results on Clean Dataset

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 Supplementary Material. 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, which can be seen in Supplementary Material.

4.1.5 Noisy Dataset for Testing Roubustness

To evaluate the robustness of TAOTF-based models compared with other methods, we ask for ophthalmologists and conclude 13 common data corruption of fundus images and classified into 4 types. Then we simulated these possible data challenges to build a noisy test set for test model robustness performances. For example, the geometric transformation could test model performance to the position deviation, viewing angle deviation, and data size deviation, the spatial transformation could test the model performance to the change of light, color, contrast, and brightness, and finally test the model performance to fuzzy images and images with more noise. Our method improves model robustness by modifying the training process (TAOTF is a model-agnostic training framework) rather than

Experiment on Noisy APTOS 2019													
	Clean	Noise			Blur			Weather			Digital		
Methods	Clean	Gaussian.	ISO.	Multiplicative.	Gauss.	Median	Motion	Optical	Rotate	RGB	Bright	Frog	Saturation
ResNet	91.17%	87.32%	89.13%	87.05%	75.60%	63.80%	78.14%	79.86%	64.43%	90.22%	81.60%	63.53%	83.06%
ResNet+SRIP [2]	91.17%	89.58%	89.16%	87.95%	76.12%	64.70%	77.45%	81.61%	68.57%	90.85%	89.13%	66.82%	86.50%
ResNet+OCNN [26]	91.08%	88.50%	88.86%	87.08%	77.75%	67.51%	76.60%	80.71%	70.29%	90.01%	88.01%	65.73%	88.04%
ResNet+hard constraints	90.75%	83.79%	88.41%	84.06%	80.80%	69.74%	75.91%	85.99%	69.84%	91.76%	90.88%	72.52%	88.41%
ResNet+WaveCNet [16]	92.18%	89.93%	88.13%	87.89%	75.39%	69.23%	80.46%	**86.23%**	70.08%	89.95%	84.96%	66.43%	86.14%
TAOTF-ResNet (Ours)	**92.53%**	**93.30%**	**92.12%**	**92.66%**	**89.76%**	**82.84%**	**84.87%**	86.00%	**76.93%**	**92.66%**	**91.76%**	**88.38%**	**92.84%**
TAOTF-ResNet50+WaveCNet (Ours)	**93.12%**	**93.18%**	**92.29%**	**92.61%**	**90.01%**	**85.09%**	**85.21%**	**88.27%**	**77.01%**	**92.44%**	**92.37%**	**89.93%**	**92.91%**
ViT3	90.99%	90.10%	89.49%	87.23%	74.47%	63.80%	74.02%	70.40%	66.34%	89.47%	88.35%	73.10%	83.15%
ViT3+orth-initialization	89.92%	87.95%	87.91%	85.19%	75.39%	63.06%	71.92%	77.00%	63.53%	87.59%	88.25%	72.07%	80.43%
ViT3+hard constraints	87.62%	86.32%	88.04%	85.24%	75.45%	63.10%	72.61%	76.30%	69.54%	87.23%	88.92%	71.26%	81.52%
ViT3+SRIP	91.07%	90.12%	90.55%	90.19%	70.95%	68.24%	73.22%	74.68%	66.33%	91.67%	90.21%	69.56%	89.69%
ViT3+satge2 (self-attention layers)	92.60%	92.32%	93.04%	89.52%	72.92%	65.52%	76.93%	79.53%	64.49%	92.66%	93.57%	61.08%	89.61%
ViT3+satge2 (transformer blocks)	95.74%	93.39%	94.38%	93.75%	75.45%	**75.79%**	75.75%	79.31%	65.43%	93.84%	94.38%	67.30%	92.78%
ViT3+satge2 (patch embdding)	95.11%	93.84%	93.75%	94.03%	72.94%	69.11%	76.99%	80.19%	63.98%	94.47%	92.84%	61.65%	91.39%
TAOTF-ViT3 (Ours)	**95.87%**	**95.87%**	**95.81%**	**95.82%**	**86.23%**	75.39%	**80.92%**	**87.14%**	**74.94%**	**95.87%**	**95.56%**	**74.70%**	**95.29%**
ViT6	92.18%	88.89%	87.17%	88.40%	77.71%	72.06%	74.86%	78.36%	67.29%	89.88%	88.90%	75.78%	81.33%
ViT6+hard constraints	91.69%	77.66%	75.30%	80.65%	51.39%	50.33%	57.26%	62.26%	50.42%	78.11%	75.75%	52.78%	76.06%
ViT6+SRIP	93.32%	88.68%	89.58%	90.40%	71.89%	66.03%	76.39%	74.79%	66.18%	89.04%	88.77%	69.72%	85.18%
TAOTF-ViT6 (Ours)	**94.06%**	**94.05%**	**93.12%**	**93.51%**	**78.82%**	**76.30%**	**76.45%**	**79.62%**	**69.82%**	**93.30%**	**92.84%**	**76.48%**	**90.67%**
Experiment on Glaucoma Detection													
ResNet	92.97%	88.67%	90.67%	88.50%	89.50%	87.50%	90.33%	89.00%	75.17%	89.67%	90.17%	76.50%	74.00%
ResNet+SRIP	94.00%	92.50%	90.50%	89.00%	89.00%	90.50%	89.00%	86.00%	77.67%	88.67%	89.50%	82.17%	80.00%
ResNet+OCNN	93.83%	87.50%	85.50%	88.50%	88.50%	88.50%	88.50%	87.17%	75.67%	85.67%	88.80%	82.17%	74.67%
ResNet+hard constraints	92.50%	91.00%	89.00%	90.17%	91.00%	86.00%	90.37%	87.50%	78.33%	90.50%	90.17%	86.70%	81.97%
TAOTF-ResNet (Ours)	**94.67%**	**94.50%**	**94.00%**	**93.97%**	**93.92%**	**93.17%**	**93.67%**	**93.77%**	**84.00%**	**93.97%**	**93.67%**	**91.09%**	**92.84%**
Experiment on Skin Lesion Classification													
ResNet50	92.89%	87.25%	86.21%	86.21%	87.17%	86.38%	86.54%	87.25%	59.62%	84.48%	81.94%	81.93%	79.93%
ResNet50+OCNN	92.74%	85.27%	86.29%	86.39%	87.12%	85.24%	86.71%	85.93%	59.85%	84.92%	82.36%	82.42%	78.23%
ResNet50+WaveCNet	90.04%	88.00%	88.06%	87.88%	88.07%	87.99%	85.41%	86.84%	61.86%	85.49%	83.21%	80.07%	81.00%
ResNet50+SRIP	92.53%	86.24%	85.45%	84.38%	86.80%	86.24%	84.06%	86.23%	57.75%	83.99%	82.57%	80.69%	78.38%
ResNet50+hard constraints	91.46%	87.97%	87.34%	89.14%	89.17%	88.42%	88.62%	88.02%	61.69%	86.21%	86.26%	84.75%	82.27%
TAOTF-ResNet50 (Ours)	**94.08%**	**94.06%**	**92.26%**	**94.02%**	**93.69%**	**92.63%**	**91.72%**	**94.36%**	**63.54%**	**91.29%**	**90.68%**	**88.12%**	**86.95%**

Table 1: Results on medical test sets (Classification). We mainly used the ViT3 model to conduct ablation experiments on the noisy Kaggle APTOS 2019, and the experiment results confirmed the role of each component of TAOTF. “orth-initialization" means that we select orthogonal initialization weights for training. “hard constraints" means that we impose hard orthogonal constraints (retraction-based manifold optimization algorithm) on model layers. For compared methods, we set the same hyperparameters for a fair comparison.

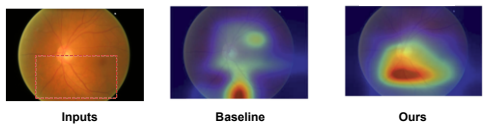

Figure 4: Glaucoma Detection Classification visualization by class activation maps (CAM) [34]. Our framework can help models find the accurate location of lesions (key features).

modifying the model structure (such as WaveCNet), these two types of methods are complementary and can be used together to enhance model robustness.

4.1.6 Experimental Results on Noisy Data

We also have conducted ablation experiments for each part of the framework based on the ViT3 model. The experimental results show that all parts of our framework have improved the robustness of the model to a certain extent. And the results of the test set are summarized in Tab. 1. It can be seen that the proposed TAOTF-based models have better robustness performances than other existing methods in this task. Because proper orthogonality can generate a more uniform spectrum of W and makes network layers closer to a 1-Lipschitz function. Given a slight perturbation to the input Δx, the variation of the output Δy is bounded low, producing a robust and less sensitive representation of data perturbations.

4.2 Experiments on Glaucoma Detection Dataset

4.2.1 Experimental Setup

For datasets, we choose a Kaggle Glaucoma Detection Dataset to test our framework performance. The dataset contains 650 images/OCT scans of the eyes. The labels were provided by clinicians who rated the Glaucoma in each image on a scale of “0, 1", which means “*No Glaucoma*", and “*Glaucoma*" respectively. We randomly shuffle the entire dataset into three subgroups, *i.e.*, training (70%), validation (10%), and testing (20%).

For the training process, in the experiments, we choose the ResNet18 model to classify Glaucoma and test the performance of our framework. We use the Ranger with $lr = 3 \times 10^{-5}$, with weight decay 1e-3, and train it for 130 epochs. The weight λ of the regularization loss is 10^{-3}. In the above training process, we use CE Loss as the criterion.

4.2.2 Experimental Results

We compare our TAOTF with prior works. And our framework significantly outperforms existing methods, which shows that TAOTF-based models have stronger robustness in this medical dataset. See Tab. 1 for a detailed comparison. After that, our model visualization can be seen Fig. 3 and Fig. 4.

4.3 Experiments on Skin Lesion Classification

4.3.1 Experiment Setup

For datasets, this dataset contains the training data for the ISIC 2019 challenge [5], and datasets from previous years (2018 and 2017).

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%**
Experiment on CIFAR-10													
MobileViT	83.53%	80.03%	72.30%	74.75%	69.21%	69.37%	74.73%	80.59%	69.15%	81.50%	81.90%	83.06%	81.18%
MobileViT+SRIP	**84.35%**	79.61%	71.99%	74.29%	69.24%	69.30%	75.14%	80.37%	69.72%	81.64%	81.89%	83.27%	81.26%
MobileViT+hard constraints	77.80%	77.19%	69.88%	71.64%	71.44%	70.44%	76.42%	80.83%	70.95%	81.28%	82.51%	83.28%	81.00%
TAOTF-MobileViT (Ours)	84.10%	**81.44%**	**75.33%**	**77.94%**	**75.10%**	**74.32%**	**77.72%**	**82.20%**	**72.35%**	**83.49%**	**84.97%**	**84.57%**	**83.17%**
VGG19 [23]	92.69%	90.43%	85.06%	85.00%	37.59%	56.12%	72.78%	85.39%	81.59%	91.74%	91.77%	92.69%	90.72%
VGG19+hard constraints	88.51%	85.54%	79.95%	79.99%	37.78%	50.43%	67.11%	78.13%	75.95%	85.55%	86.51%	88.51%	85.14%
VGG19+SRIP	93.08%	90.34%	85.74%	86.12%	39.08%	55.08%	73.68%	85.13%	82.74%	91.71%	91.95%	93.04%	91.10%
TAOTF-VGG19 (Ours)	**93.70%**	**93.11%**	**92.51%**	**89.67%**	**89.08%**	**45.27%**	**59.28%**	**77.21%**	**85.64%**	**86.29%**	**92.21%**	**91.76%**	**92.73%**

Table 2: Results on CIFAR-100/CIFAR-10 test sets. Especially, to demonstrate the generalization performance of TAOTF for various neural network models. We conducted as many model tests as possible and simulated 12 common data challenges for testing. In order to compare with other methods and control variables more strictly, we selected the same basic settings of the experiment for a fair comparison, without taking complex data enhancement methods and other tricks.

[24] [9]. The dataset contains 25331 images available to classify dermoscopic images among nine diagnostic categories. The labels were provided by clinicians who rated the classification of a skin lesion in each image on a scale of "0, 1, 2, 3, 4, 5, 6, 7, 8", which means "*Melanoma*", "*Melanocytic nevus*", "*Basal cell carcinoma*", "*Actinic keratosis*", "*Benign keratosis*", "*Dermatofibroma*", "*Vascular lesion*", "*Squamous cell carcinoma*", "*None of the above*". We randomly shuffled the entire dataset into three subgroups, *i.e.*, training (80%), validation (10%), and testing (10%).

For training, we choose ResNet-50 to test our framework [17]. We use the Adam with $lr = 0.001$, and train it for 100 epochs. The weight λ of the regularization loss is 10^{-4}. In the above training, we use the CE Loss as the criterion [18].

4.3.2 Experimental Results

We compare our framework TAOTF with other methods. See Tab. 1 for a detailed comparison. The above results confirm that imposing proper orthogonality constraints for models has stronger robustness performances on this noisy test dataset.

4.4 Experiments on Brain MRI segmentation

4.4.1 Experiment Setup

The dataset contains brain MR images together with manual FLAIR abnormality segmentation masks. The dataset containing 3929 images was obtained from The Cancer Imaging Archive (TCIA). They correspond to 110 patients included in The Cancer Genome Atlas (TCGA) lower-grade glioma collection with at least fluid-attenuated inversion recovery (FLAIR) sequence and genomic cluster data available. We randomly shuffle the entire dataset into three subgroups, *i.e.*, training (70%), validation (10%), and testing (20%). We choose UNet [22], which is composed of 10 convolution layers. Training takes 30 epochs with the TAOTF regularizer applied to the model. The weight λ of the regularization loss is 10^{-3} and all of other settings retain as standard default. In the above training process, we use Binary Cross Entropy (BCE) and Dice loss as the criterion.

4.4.2 Experiment Results

We compare TAOTF-based UNet with other methods. See Tab. 3 for a detailed comparison. The above results confirm that TAOTF-based models can help improve segmentation tasks. For lambda, we mainly use the parameter from as default [2] . We test the model performance in different lambada settings in stage 2 and find that the choice of lambda has little impact within the above range(10^{-5}, 10^{-3}). Part of the reason is that the first stage has already found a good starting point on the Stiefel manifold (strong orthogonal weight matrices).

	BCE (↓)	**ACC (↑)**	**F1-Score (↑)**
UNet	0.00713	0.996	0.837
UNet+SRIP	0.00949	0.996	0.811
TAOTF-UNet (Ours)	**0.00697**	**0.997**	**0.843**

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

4.5 Experiments on CIFAR-10/CIFAR-100

4.5.1 Experiment Setup

For datasets, the CIFAR-10 dataset has a total of 60000 color images. These images are $32 \times 32 \times 3$ and are divided into 10 categories, with 6000 images in each category. Among them, 50000 images are used for training, another 10000 images are used for testing. The CIFAR-100 dataset has 100 categories, with 500 training images and 100 testing images per category.

For training, in the first dataset CIFAR-10, we choose the MobileViT-s [20], which combined with the transformers and CNNs, to test our framework performance on the CIFAR-10 dataset. We use the AdamW with $lr = 0.001$ to train it for 60 epochs. In the second dataset, we choose 28-depth WideResNet [28] to test our framework performance on the CIFAR-100 dataset further. We use the Adam with $lr = 0.0005$ to train it for 80 epochs. The weight λ of the regularization loss is 10^{-4}. In the above training, the criterion chosen is CE Loss with 0.1 label smoothing to reduce overfitting.

4.5.2 Experimental Results

In order to demonstrate the generalization performance of TAOTF, we only select the most representative models to show generalization performance of TAOTF, such as wideResNet, a representative of increasing the width of networks, and the mobileViT, which combines transformers and CNNs. We compare our framework TAOTF with other methods. See Tab. 2 for a detailed comparison. The above

results confirm that TAOTF-based models can also resist the corruption of natural images to a certain extent.

To better understand how our framework works, we design a simple auxiliary denoising experiment on CIFAR-10. We add random noise with noise factor 0.1 and choose a simple denoising model DnCNN [32] with the corresponding depth of 16 to test our framework. We use the Adam with $lr = 0.001, batchsize = 32$, and train it for 35 epochs. The weight λ of the regularization loss is 10^{-4}. In the above training, we use the MSE Loss as the criterion. See Tab. 4 for a detailed comparison. Our TAOTF-based models can ignore small input perturbations (*e.g.*, noise, blur), enjoying robustness under noisy data.

	PSNR (↑)	**SSIM (↑)**	**LPIPS (↓)**
CNN	26.697	0.893	0.092
TAOTF-CNN (Ours)	**29.117**	**0.942**	**0.084**

Table 4: Results on Denoising CIFAR-10 test sets. "PSNR" is Peak Signal to Noise Ratio, "SSIM" is Structural Similarity, and "LPIPS" is Learned Perceptual Image Patch Similarity [33]. Proper orthogonality constraints can help the model ignore small input perturbations and improve model performance facing low-level tasks (*e.g.*, noise, blur).

5 Conclusion

In this paper, according to the practical difficulties encountered in data quality, we proposed a new two-stage training framework TAOTF, which can find a trade-off between the orthogonality constraint and the main task solution set, and propose an orthogonal initialization algorithm PDOI in the first stage that can find a suitable starting point for orthogonal optimization. Then we use the gradient descent algorithm with soft orthogonal constraints to find a better point. Our framework was tested both on transformers and CNNs, and the experimental results show that our framework can significantly improve the robustness performances of these models facing noisy data.

6 Acknowledgement

This work was supported by the National Natural Science Foundation of China (No. 62101351) and the GuangDong Basic and Applied Basic Research Foundation (No.2020A1515110376).

References

[1] P-A Absil, Robert Mahony, and Rodolphe Sepulchre, 'Optimization algorithms on matrix manifolds', in *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, (2009).

[2] Nitin Bansal, Xiaohan Chen, and Zhangyang Wang, 'Can we gain more from orthogonality regularizations in training deep networks?', *Advances in Neural Information Processing Systems*, **31**, (2018).

[3] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin, 'Albumentations: Fast and flexible image augmentations', *Information*, **11**(2), (2020).

[4] Jie Chen and Yousef Saad, 'On the tensor svd and the optimal low rank orthogonal approximation of tensors', *SIAM journal on Matrix Analysis and Applications*, **30**(4), 1709–1734, (2009).

[5] Marc Combalia, Noel CF Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C Halpern, Susana Puig, et al., 'Bcn20000: Dermoscopic lesions in the wild', *arXiv preprint arXiv:1908.02288*, (2019).

[6] Yanhong Fei, Yingjie Liu, Xian Wei, and Mingsong Chen, 'O-vit: Orthogonal vision transformer', *arXiv e-prints*, (2022).

[7] Shiv Gehlot, Anubha Gupta, and Ritu Gupta, 'A cnn-based unified framework utilizing projection loss in unison with label noise handling for multiple myeloma cancer diagnosis', *Medical Image Analysis*, 102099, (2021).

[8] Chengyue Gong, Dilin Wang, Meng Li, Vikas Chandra, and Qiang Liu, 'Vision transformers with patch diversification', *arXiv preprint arXiv:2104.12753*, (2021).

[9] David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern, 'Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic)', *arXiv preprint arXiv:1605.01397*, (2016).

[10] Mehrtash Harandi and Basura Fernando, 'Generalized backpropagation,\'{E} tude de cas: Orthogonality', *arXiv preprint arXiv:1611.05927*, (2016).

[11] Wei Hu, Lechao Xiao, and Jeffrey Pennington, 'Provable benefit of orthogonal initialization in optimizing deep linear networks', *arXiv preprint arXiv:2001.05992*, (2020).

[12] Lei Huang, Xianglong Liu, Bo Lang, Adams Yu, Yongliang Wang, and Bo Li, 'Orthogonal weight normalization: Solution to optimization over multiple dependent stiefel manifolds in deep neural networks', in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, (2018).

[13] Zhiwu Huang and Luc Van Gool, 'A riemannian network for spd matrix learning', in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, (2017).

[14] Mario Lezcano-Casado and David Martınez-Rubio, 'Cheap orthogonal constraints in neural networks: A simple parametrization of the orthogonal and unitary group', in *International Conference on Machine Learning (ICML)*, pp. 3794–3803. PMLR, (2019).

[15] Jun Li, Fuxin Li, and Sinisa Todorovic, 'Efficient riemannian optimization on the stiefel manifold via the cayley transform', in *International Conference on Learning Representations (ICLR)*, (2019).

[16] Qiufu Li, Linlin Shen, Sheng Guo, and Zhihui Lai, 'Wavecnet: Wavelet integrated cnns to suppress aliasing effect for noise-robust image classification', *IEEE Transactions on Image Processing*, **30**, 7074–7089, (2021).

[17] Lei Liu, Wentao Lei, Xiang Wan, Li Liu, Yongfang Luo, and Cheng Feng, 'Semi-supervised active learning for covid-19 lung ultrasound multi-symptom classification', in *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, p. 1268–1273. IEEE, (2020).

[18] Li Liu, Gang Feng, Denis Beautemps, and Xiao-Ping Zhang, 'A novel resynchronization procedure for hand-lips fusion applied to continuous french cued speech recognition', in *2019 27th European Signal Processing Conference (EUSIPCO)*, p. 1–5. IEEE, (2019).

[19] Li Liu, Thomas Hueber, Gang Feng, and Denis Beautemps, 'Visual recognition of continuous cued speech using a tandem cnn-hmm approach.', in *Interspeech*, p. 2643–2647, (2018).

[20] Sachin Mehta and Mohammad Rastegari, 'Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer', in *International Conference on Learning Representations*, (2021).

[21] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio, 'On the difficulty of training recurrent neural networks', in *International conference on machine learning (ICML)*, pp. 1310–1318. PMLR, (2013).

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, 'U-net: Convolutional networks for biomedical image segmentation', in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, (2015).

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

[24] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler, 'The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions', *Scientific data*, **5**(1), 1–9, (2018).

[25] Jianrong Wang, Ziyue Tang, Xuewei Li, Mei Yu, Qiang Fang, and Li Liu, 'Cross-modal knowledge distillation method for automatic cued speech recognition', 2986–2990, (2021).

[26] Jiayun Wang, Yubei Chen, Rudrasis Chakraborty, and Stella X Yu, 'Orthogonal convolutional neural networks', in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11505–11515, (2020).

[27] Di Xie, Jiang Xiong, and Shiliang Pu, 'All you need is beyond a good init: Exploring better solution for training extremely deep convolutional neural networks with orthonormality and modulation', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6176–6185, (2017).

[28] Sergey Zagoruyko and Nikos Komodakis, 'Wide residual networks', in *British Machine Vision Conference 2016*. British Machine Vision Association, (2016).

[29] Aston Zhang, Alvin Chan, Yi Tay, Jie Fu, Shuohang Wang, Shuai Zhang, Huajie Shao, Shuochao Yao, and Roy Ka-Wei Lee, 'On orthogonality constraints for transformers', in *ACL/IJCNLP (2)*, (2021).

[30] Chunhui Zhang, Guanjie Huang, Li Liu, Shan Huang, Yinan Yang, Xiang Wan, Shiming Ge, and Dacheng Tao, 'Webuav-3 m: A benchmark for unveiling the power of million-scale deep uav tracking', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2022).

[31] Chunhui Zhang, Li Liu, Yawen Cui, Guanjie Huang, Weilin Lin, Yiqian Yang, and Yuehong Hu, 'A comprehensive survey on segment anything model for vision and beyond', *arXiv preprint arXiv:2305.08196*, (2023).

[32] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang, 'Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising', *IEEE transactions on image processing*, **26**(7), 3142–3155, (2017).

[33] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, 'The unreasonable effectiveness of deep features as a perceptual metric', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, (2018).

[34] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, 'Learning deep features for discriminative localization', in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2921–2929, (2016).