Towards an Effective Orthogonal Dictionary Convolution Strategy

Yishi Li [1,2], **Kunran Xu** [1,2], **Rui Lai** [1,2*], **Lin Gu** [3,4*]

[1] School of Microelectronics, Xidian University, Xi'an 710071, China
[2] Chongqing Innovation Research Institute of Integrated Cirtuits, Xidian University, Chongqing 400031, China
[3] RIKEN AIP, Tokyo 103-0027, Japan
[4] The University of Tokyo, Tokyo, Japan
yshlee1994@outlook.com, aazzttcc@gmail.com, rlai@mail.xidian.edu.cn, lin.gu@riken.jp

Abstract

Orthogonality regularization has proven effective in improving the precision, convergence speed and the training stability of CNNs. Here, we propose a novel Orthogonal Dictionary Convolution Strategy (ODCS) on CNNs to improve orthogonality effect by optimizing the network architecture and changing the regularized object. Specifically, we remove the nonlinear layer in typical convolution block "Conv(BN) + Nonlinear + Pointwise Conv(BN)", and only impose orthogonal regularization on the front Conv. The structure, "Conv(BN) + Pointwise Conv(BN)", is then equivalent to a pair of *dictionary* and *encoding*, defined in sparse dictionary learning. Thanks to the exact and efficient representation of signal with dictionaries in low-dimensional projections, our strategy could reduce the superfluous information in dictionary Conv kernels. Meanwhile, the proposed strategy relieves the too strict orthogonality regularization in training, which makes hyper-parameters tuning of model to be more flexible. In addition, our ODCS can modify the state-of-the-art models easily without any extra consumption in inference phase. We evaluate it on a variety of CNNs in small-scale (CIFAR), large-scale (ImageNet) and fine-grained (CUB-200-2011) image classification tasks, respectively. The experimental results show that our method achieve a stable and superior improvement.

Introduction

With the development of deep learning research, convolutional neural networks (CNNs) (Krizhevsky, Sutskever, and Hinton 2012; Simonyan and Zisserman 2015; He et al. 2016) have been increasingly demonstrated to offer efficient feature extraction capabilities. However, the booming parameters and complex structures make it easy to incur vanishing/exploding gradients (Bengio, Simard, and Frasconi 1994; Glorot and Bengio 2010), especially for ultra-deep models. To constrain these huge parameters in high-dimension spaces, many solutions have been invented, including parameter initialization (Saxe, McClelland, and Ganguli 2014), normalization of internal activation (Ioffe and Szegedy 2015), second-order optimization (Dauphin et al. 2014), *et al*.

*Corresponding author.

Among these methods, orthogonality regularization (Xie, Xiong, and Pu 2017) is a commonly used training technique to be applied on the linear transformations on hidden layers of CNNs. By forcing filters in convolution (Conv) kernel to be orthogonal from each other, orthogonality regularization stabilizes the distribution of each hidden layer's activation, reduces the phenomenon of gradient vanishing or exploding and accelerates the convergence speed (Zhou, Do, and Kovacevic 2006). More recently, a lot of variations (Wang et al. 2020; Yang et al. 2020; Sarhan et al. 2020; Liu et al. 2019) that further reduce the correlation of each kernel are successively presented, thus improving the feature expressiveness, robustness, and the performances in tasks.

Inspired by the success of sparse dictionary learning (Mailhé et al. 2008; Yaghoobi, Blumensath, and Davies 2008; Mairal et al. 2009; Ramírez, Sprechmann, and Sapiro 2010), we take notice of the similarity between Conv block and sparse coding methods (Olshausen and Field). We observe that Convs play a similar role as dictionaries that save large prior information. Since reducing correlation between filters in kernel has proven to be benefit for performance (Rabouy, Paris, and Glotin 2015), we further speculate that it is better to impose orthogonality regularization on *dictionary* layers than on all of them to maximize the effect of orthogonality. Therefore, we propose our Orthogonal Dictionary Convolution Strategy (ODCS) that enhances the effect of orthogonality by optimizing the structure of CNNs. As shown in Figure 1, given a typical Conv block structure, our ODCS has three main steps: (1) Create or locate the specific structure of "Conv(BN) + Nonlinear + Pointwise Conv(BN)" in networks. (2) Remove nonlinear layer between Convs. (3) Only regularize the kernel of front Conv by orthogonality regularization without constraint on norm of kernels, forming the pair of *dictionary* and *encoding* matrix kernel, to improve the CNNs' performance.

Our proposed strategy enhances the ability of extracting various features, and thereby fully utilizing the model capacity. Due to the removal of nonlinear function, two Convs can be fused as an extensive linear transformer or a stronger Conv. In terms of sparse dictionary learning, the former is a *dictionary* matrix and the latter is actually an *encoding* matrix. Imposing orthogonality regularization on dictionary Conv can reduce redundancy of kernels in whole structure, thus producing gain in fusion of models with orthogonal rep-

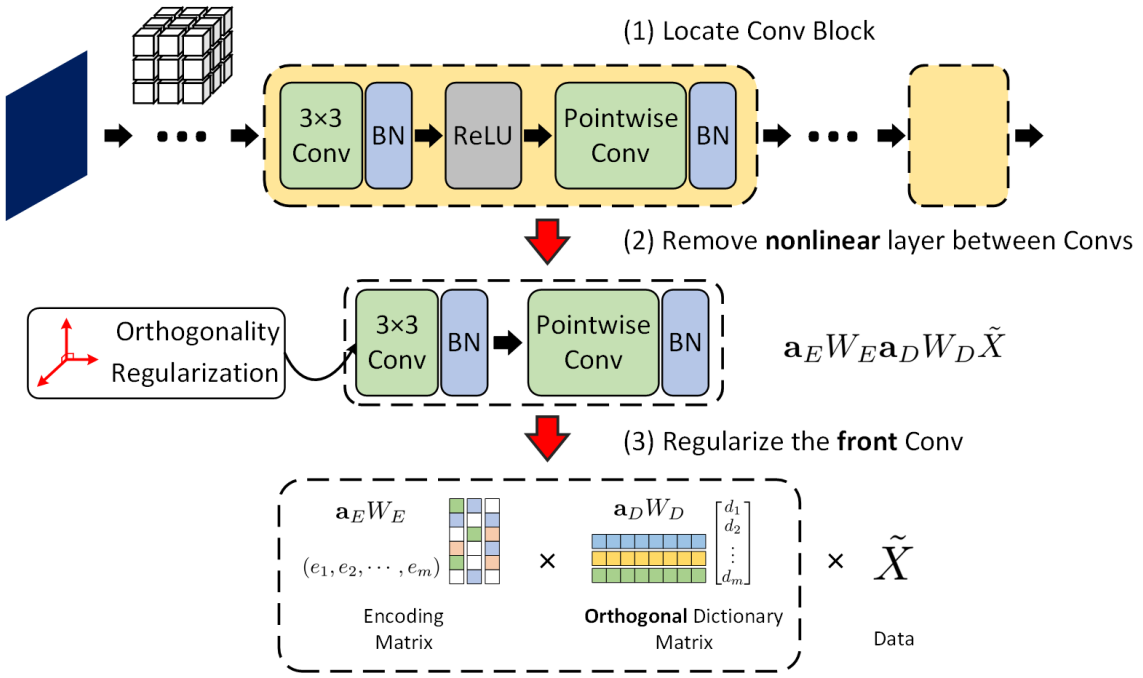

Figure 1: We propose a novel Orthogonal Dictionary Convolution Strategy (ODCS). The strategy recommends a method to optimize the structure for orthogonality regularization to maximize the effect of it. There are three main steps: (1) Locate Conv block, "Conv(BN) + Nonlinear + Pointwise Conv(BN)". (2) Remove nonlinear layer between Convs. (3) Regularize only on the front Conv. Convs is equivalent to the product of dictionary matrix and encoding matrix. Our strategy can be applied on state-of-the-art models easily without any extra consumption in inference phase.

resentations (Rabouy, Paris, and Glotin 2015). By imposing the regularization selectively and releasing the constraint on norm of kernels in traditional regularization, our ODCS relieves the negative effect of too strict orthogonality (Bansal, Chen, and Wang 2018). Our method could save the efforts for searching the optimal hyper-parameter on the weight of regularization during training. The experiments demonstrate that our ODCS yields superior performance compared to state-of-the-art orthogonality strategies.

In summary, our contributions could be summarized as:

- We propose a novel Orthogonal Dictionary Convolution Strategy (ODCS) that only imposes orthogonality regularization on dictionary Convs instead of all layers, and the regularization is without norm constraint on kernels.
- The presented strategy enhances the diversity of dictionary Conv without extra cost. Moreover, our strategy could easily be applied on the classical CNN architectures to improve the performance.
- The method relieves negative effect of too strict regularization. This makes it easier to tune the hyper-parameter setting on regularization, thus facilitating its application on different datasets and networks.
- Extensive experimental results show that the proposed ODCS consistently improves the classification accuracy in small-scale (CIFAR), large-scale (ImageNet-2012) and fine-grained (CUB-200-2011) image recognition compared to traditional methods.

Related Works

Orthogonality was used in the initialization method of CNNs early. (Saxe, McClelland, and Ganguli 2014; Mishkin and Matas 2016) exhibited random orthogonal initial conditions on network weights. The initial conditions lead to efficient propagation of gradients even in deep nonlinear networks. The initialization method allows learning of very deep networks via standard SGD to converge fast and shows the superior performance than standard initialization.

Many works focus on exploring loss function for orthogonality regularization, and it is more widely used to stabilize training. (Rodríguez et al. 2017) pointed out that feature decorrelation is an alternative for using the full capacity of the models. They imposed constraints in feature decorrelation to eliminate interference between negatively correlated feature weights to reduce over-fitting efficiently. Works (Xie, Xiong, and Pu 2017) proposed a variant of regularization that utilizes orthogonality among different filter banks without any shortcuts/identity mappings from scratch. Other works (Huang et al. 2018; Bansal, Chen, and Wang 2018) explored a variety of orthogonality regularization loss, and they proposed Spectral Restricted Isometry Property Regularization (SRIP) has better performance improvement. They verified the performance of each regularization to alleviate the gradient vanishing or exploding phenomenon in training networks. In recent works, (Wang et al. 2020) did not use the common kernel orthogonality. They proved that the orthogonality of the kernel cannot guarantee the orthog-

onality of the convolutional layer. Even if the kernel matrix satisfies the orthogonality, the Conv itself is still non-uniform and changeable. They imposed orthogonality between filters based on the doubly block-Toeplitz matrix representation of the convolutional kernel.

Many works explored other training methods to gain orthogonality of layers equivalently. Works (Harandi and Fernando 2016; Ozay and Okatani 2016; Huang et al. 2018) considered Stiefel manifold-based hard constraints of weights. They proposed several orthogonal weight normalization methods to solve optimization over multiple dependent Stiefel manifolds. MHH-based methods (Liu et al. 2018; Lin et al. 2020) are inspired by the Thomson problem in physics and define the hyperspherical energy to characterize the diversity on a unit hypersphere and shows significant and consistent improvement in supervised learning tasks. Since the orthogonal regularization is too limiting (Miyato et al. 2018), (Brock, Donahue, and Simonyan 2019) explored several variants designed to relax the constraint. They minimized the pairwise cosine similarity between filters by removing the diagonal terms from the regularization and freeing the norm.

On the other side, the sparse dictionary learning (Mailhé et al. 2008; Yaghoobi, Blumensath, and Davies 2008) has been found useful on diversity-based regularization in CNN training. Early studies in sparse coding (Mairal et al. 2009; Ramírez, Sprechmann, and Sapiro 2010), model the diversity with the empirical covariance matrix and show that encouraging such diversity can improve the dictionary's generalizability. (Rabouy, Paris, and Glotin 2015) improved image classification by orthogonality of sparse codes. To explain the observations of the networks' fusion results, they studied the orthogonality properties by the cosine computation and put forward the various qualities of the studied bases and sparse representation. Analogously, in the Low-Rank methods, (Yang et al. 2020) ensured the valid form of SVD training by adding regularization on singular vectors of each Conv.

Different from the existing works, our ODCS aims at exploring the appropriate method to utilize orthogonality regularization in sparse dictionary learning by researching the suitable structure of networks.

Method

In this section, we will review the existing classic orthogonality regularization widely used in CNN training. Following that, we will further describe our ODCS and give the proof of mathematical expression as well as corresponding analysis.

The default mathematical expressions are defined as follows. Let $W \in \mathbb{R}^{m\times n}$ donates the Conv kernel, where $m = N_k$ and $n = C_d H_k W_k$. C_d is the channel of input data. N_k, H_k and W_k are the numbers, height and width of the kernel.

Preliminaries

Previous work, Soft Orthogonality Regularization (SO), recommended that the Gram matrix of Conv kernel $W^T W$ should be approximated to the identity matrix. The regularization is implemented as:

$$\lambda \sum_{W} ||W^T W - I||_2^2 \tag{1}$$

where I donates the identity matrix and $||\cdot||_2$ is *l2-norm*. λ indicates the weight of loss.

This regularization has two implicit meanings to filters in kernel. It constrains convolution filters to be orthogonal to each other, and enforces the modulus norm of each filter to be consistent with 1. It takes advantage of orthogonality while maintains the normal propagation of the gradient during training. The strictness of regular constraints can be controlled simply by adjusting λ. There are many variants of orthogonality regularization with different loss functions and training strategies.

The Proposed Orthogonal Dictionary Convolution Strategy

The structure "Conv(BN) + Nonlinear + Pointwise Conv(BN)" frequently applied in recent networks, like ResNet (He et al. 2016) and its variants (Xie et al. 2017; Zagoruyko and Komodakis 2016) and DenseNet (Huang et al. 2017). The front Conv plays the main role of extracting spatial features and the Pointwise Conv performs dimensionality reduction and expansion (He et al. 2016). Inspired by sparse dictionary learning, during the inference, it's seem like the front Conv plays the role of *dictionary* and the Pointwise one is like *encoding*. Meanwhile, orthogonal dictionary is proved to enhance the performance of feature extractor (Rabouy, Paris, and Glotin 2015). Therefore, we explore to apply the sparse dictionary learning upon CNNs from the mathematical formula and abundant experiments.

Formulas for ODCS. With **im2col** method (Heide, Heidrich, and Wetzstein 2015; Yanai, Tanno, and Okamoto 2016), kernel W is retained and data X is converted to patch-matrix $\tilde{X}$. $\tilde{X} \in \mathbb{R}^{n\times k}$, where $k = W_d H_d$. W_d and H_d are width and height of data. Then, Conv can be formulated as matrix multiplications of W and $\tilde{X}$. The result Y calculated by the unit consisting of Conv, batch normalization (BN), and ReLU could be denoted as

$$\begin{aligned} Y &= \psi(\gamma \frac{W\tilde{X} - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta) \\ &= \psi(\gamma \frac{W\tilde{X}}{\sqrt{\sigma^2 + \epsilon}} + \beta - \gamma \frac{\mu}{\sqrt{\sigma^2 + \epsilon}}) \\ &= \psi(\mathbf{a} W\tilde{X} + \mathbf{b}) \end{aligned} \tag{2}$$

where μ, σ^2, γ, β are the parameters of BN and ψ denotes the ReLU. In order to simplify the equation, let $\mathbf{a} = \gamma/\sqrt{\sigma^2 + \epsilon}$ and $\mathbf{b} = \beta - \mathbf{a}\mu$. The structure, "Conv(BN) + Nonlinear + Pointwise Conv(BN)", can be expressed as

$$Y = \mathbf{a}_E W_E \psi(\mathbf{a}_D W_D \tilde{X} + \mathbf{b}_D) + \mathbf{b}_E \tag{3}$$

where W_D and W_E are the kernels of front Conv and latter Pointwise Conv respectively. $\mathbf{a}_D$, $\mathbf{b}_D$, $\mathbf{a}_E$, $\mathbf{b}_E$ are parameters of BN after each Conv. There are two points worth

noted. Firstly, when the kernel size of Conv is 1 (Pointwise), the conversion of X can be ignored, as $\tilde{X} = X$. Secondly, if **remove ReLU** ψ between two Convs, the equation then is changed to

$$Y = \mathbf{a}_E W_E \mathbf{a}_D W_D \tilde{X} + \mathbf{a}_E W_E \mathbf{b}_D + \mathbf{b}_E \tag{4}$$

We want to emphasize that **removing the nonlinear layer** and **insuring the latter Conv to be Pointwise** are two vital steps in our ODCS. Our ODCS applies the theory of sparse dictionary learning in CNN training. According to the theory, linearity of blocks is the key to apply the theory in CNN successfully. This is also empirically validated by our experiments.

Without nonlinear operation, two multipliers, $\mathbf{a}_D W_D$ and $\mathbf{a}_E W_E$, is strictly combined to the linear transformation. We have argued above that these two Convs are functionally different. Specifically, with the view of sparse dictionary learning (Olshausen and Field; Lee et al. 2006; Rabouy, Paris, and Glotin 2015), every row in $\mathbf{a}_D W_D$ can be regarded as a "basis" kernel and every column in $\mathbf{a}_E W_E$ is a set of "coefficient" vectors. The linear combination of kernels is equivalent to a more complex kernel, like the signal recovered. In view of this, we call the W_D as dictionary Conv and W_E encoding Conv.

Orthogonal dictionary Conv. According to the analysis in (Bansal, Chen, and Wang 2018), if the kernel is over-complete, reducing the correlation within dictionary kernel would increase the diversities. In prevalent deep learning networks, W_D is always an over-complete matrix. Therefore, the orthogonality on W_D is useful in most CNNs. On the other side, in works (Rabouy, Paris, and Glotin 2015; Tropp and Gilbert 2007) in sparse coding, cosine computation is used to improve the orthogonality properties, gaining the performance improvement in image classification. Diversity in dictionary in sparse methods has been proved to be beneficial. Inspired by these conclusions, in order to improve the feature extraction capability, we propose to impose orthogonality regularization on dictionary Convs as follows:

$$L_{ODCS} = \lambda \sum_{W \in \mathbb{W}_D} ||W^T W \circ (\mathbf{1} - I)||_2^2 \tag{5}$$

where $\mathbb{W}_D$ denotes the set of dictionary Convs according to ODCS in the network. $\circ$ denotes the element-wise multiplication and $\mathbf{1}$ is the matrix with all elements set to 1. Comparing Equation 5 and Equation 1, it is clear there are two significant differences: (1) We relax the constraint on the norm. The norm of kernels of dictionary Conv can be trained freely when the kernel is upgraded. (2) We only constrain the directions of each filter as suggested by (Brock, Donahue, and Simonyan 2019) while Equation 1 further limits the norm of all kernels to 1. (Bansal, Chen, and Wang 2018) has proven that the over strict regularization like Equation 1 may harm the model capacity and expressiveness. We will experimentally demonstrate that our would alleviate the side effect of existing orthogonality regularization.

Finally, we add the regularization loss with weight λ to the final loss for CNNs, so that the task objective and orthogonality regularization can be simultaneously achieved.

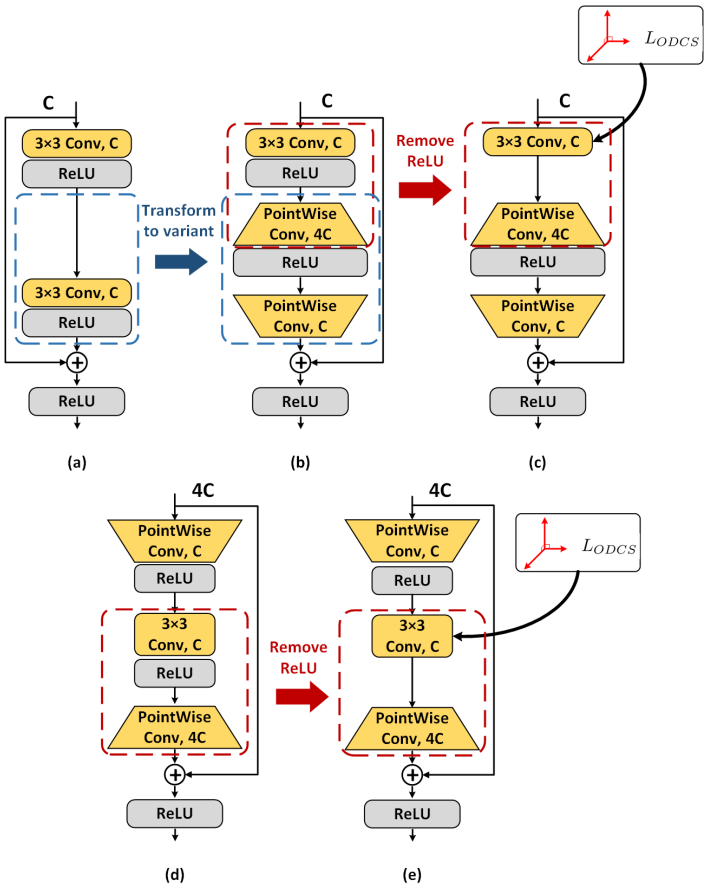

Figure 2: ODCS applied on different networks. (a) Basic block in ResNet; (b) Variant of (a). The second 3×3 Conv is decomposed into two Pointwise Convs to fit the premise of ODCS; (c) The structure modified from (b) by ODCS; (d) BottleNeck block in ResNet; (e) The structure modified from (d) by ODCS. In the all the structures, BN layers are omitted. The red boxes in (c) and (e) mark the structure we proposed.

In addition, as shown in Equation 4, the regularization should strictly constrain the product of Conv kernel and parameters of BN, $\mathbf{a}_D W_D$, instead of W_D. In fact, they are essentially the same. Although γ and σ would change for each batch of training data, they could be also regarded as parameters and be merged into Conv. In addition, since the $\mathbf{a}_D$ only alter the norm rather than the direction of every vector in the matrix W_D, the orthogonality of W_D would not be affected by BN. To simplify, we just impose orthogonality regularization on the kernel of dictionary Conv.

Applications of ODCS on variety CNNs.

Our strategy is suitable for most classical network backbones. In order to adapt the backbone to our proposed ODCS, we will show more examples. We need to redesign some blocks and apply our ODCS.

Blocks without Pointwise Conv. Basic block in ResNet

is taken as represent of plain networks without Pointwise Conv, as shown in figure 2(a). The original basic block used in ResNet for small-scale images consists of two 3×3 Convs and does not include Pointwise Conv, so it is necessary to be introduced.

Since the amount of parameters and calculations effect the performance of the network heavily, we redesign the basic block based on the purposes that maintaining the parameters and calculations. Meanwhile, our ODCS needs a Pointwise Conv behind the 3×3 Conv. We decompose the second 3×3 Conv into two Pointwise Convs, and the output channel of the first Pointwise Conv is set to $4C$ (the number of input channel is C), as shown in figure 2(b). The variant has a similar structure, parameters, and calculations to the original basic block. In the experiment, we test the performance of this network adequately to prove that the variant can be the baseline of our experiment.

Then, the variant has satisfied the condition. We apply the ODCS to the variant by removing the ReLU behind 3×3 Conv and imposing the loss L_{ODCS} to the 3×3 Conv, as shown as in figure 2(c).

Blocks with Pointwise Conv. The bottleneck block, in figure 2(d), is widely used in networks. It originally has the structure "3×3 Conv BN + ReLU + Pointwise Conv BN". So, we directly apply ODCS to the bottleneck block and gain the recommended structure in Figure 2(e).

Experiments

First of all, we benchmark our Orthogonal Dictionary Convolution Strategy on ResNet (He et al. 2016) including several different variants. All training/validation data, data preprocessing process, and data enhancement process are kept consistent in ablation experiments. Top-1 accuracy is evaluated in each experiment.

Small-scale Image Classification on CIFAR

CIFAR-10 and CIFAR-100 consists of 50k training images and 10k validation images, divided into 10 and 100 classes respectively. We use the standard SGD optimizer to train our models with momentum of 0.9 and the weight decay is 4×10^{-5}. We use data augmentations, including random cropping and random flip. These models are trained with a mini-batch size of 128 on one GPU. In each experiment, we train several times with the same configuration to prevent the impact of fluctuations, and report the median of results.

We perform two groups of experiments using *basic block* and *bottleneck block* for small-scaled image classification task. In order to adapt the networks to our proposed ODCS, we transform the basic block in ResNet-20, 56 and 110 to the variant according to strategy, as shown in Figure 2(b). In order to distinguish the ResNet and its variants, we mark the original ResNet as "Origin" and mark the variants changed by ODCS(without L_{ODCS}) as "baseline" in Table 1. The results of original ResNets and variants are compared to confirm the effect of modify in Table 1. It shows that the accuracy of variants are approximate to original networks. Therefore, the structural alteration might not be the main reason for improvement of ODCS. We test the SO and SRIP (Bansal, Chen, and Wang 2018) and list the results of SVD (Yang et al. 2020) and ONI (Huang et al. 2020). It's obvious that ODCS is better than other works in absolute precision.

We also test the bottleneck block in ResNet-50 and 101 in Table 2. Compared with the work (Wang et al. 2020), our baseline is 4.4% higher. The amount of increase is 1.72% and is also larger than the latest work. Our ODCS has shown the remarkable performance gains in all experiments and improvement is stable.

	CIFAR-10			CIFAR-100		
Method	20	56	110	20	56	110
Origin	92.07	93.25	92.64	77.90	80.12	79.92
Baseline	92.49	93.37	93.40	78.84	80.27	80.73
SO	92.46	93.59	93.46	76.80	81.55	81.58
SRIP	92.65	93.89	93.96	79.20	82.33	82.84
SVD*	91.39	93.27	93.47	-	-	-
ONI*	-	-	93.44	-	-	-
ODCS	**93.13**	**94.29**	**94.55**	**79.50**	**82.65**	**83.08**

Table 1: Accuracy on CIFAR using ResNet-20, 56 and 110 with **basic block**. "*" indicates the results reported in the cited paper.

	CIFAR-10		CIFAR-100	
Method	50	101	50	101
Baseline(ResNet)*	-	-	78.50	-
OCNN*	-	-	79.50	-
Baseline(ResNet)	92.98	92.76	82.90	83.51
ODCS	**93.57**	**93.60**	**84.62**	**84.55**

Table 2: Accuracy on CIFAR using ResNet-50 and 101 with **bottleneck block**. "*" indicates the results reported in work (Wang et al. 2020).

Fine-grained Image Classification on CUB-200-2011

We conduct fine-grained image classification experiments on CUB-200-2011 bird dataset to show the performance of ODCS. The CUB-200-2011 is a most widely studied bird's classification task, with 5994 training images and 5794 test images annotated with bounding boxes from 200 wild bird species. It is one of the most competitive datasets, since there are only 30 images in each category for training. During training, we set the batch size to 72 and the initial learning rate as 0.05 with decay factor of 0.1 after every 30 epochs to train each model for 120 epochs. We use random cropping, brightness jitter and random flip data augmentations provided by standard training setting. According to the tuning schedule for λ on CIFAR, we adjust the schedule used in CUB-200-2011 proportionally.

We use ResNet-34, 50 and 101 models as baselines and we impose the ODCS on these networks. Experiments show that ODS improves the performance of fine-grained image classification. Compared to the baselines, ODCS increased

the accuracy for 2.45%, 0.51% and 0.26% for ResNet-34, 50 and 101, respectively. The improvement we have achieved is considerable in the challenging fine-grained classification.

We find that performance of SO degraded in training. Its accuracy is less than the baseline by 0.9% and 1.15% on ResNet-50 and 101. It's also shown in Figure 3(bottom) that curve of SO fluctuates severely during the whole training, probably due to the too strict constraint. The performance of SRIP is better than SO, though its curve fluctuates more vigorously than ours.

Method	34	50	101
Baseline(ResNet)	79.48	81.04	82.70
SO	80.80	80.14	81.55
SRIP	81.45	81.17	82.82
ODCS	**81.93**	**81.55**	**82.96**

Table 3: Accuracy of fine-grained image classification task in CUB-200-2011 datasets. ResNet-34 with basic block and ResNet-50 and 101 with bottleneck block are used.

Large-scale Image Classification on Imagenet-2012

To further validate the effectiveness of ODCS on large-scale image classification, we evaluate it on the ImageNet-2012 dataset (Russakovsky et al. 2015). The experimental settings are kept as below: We apply SGD with a momentum of 0.9, and a weight decay of 4e-5. The initial learning rate is 0.4 and it is adjusted following "cosine" learning schedule. It trains 120 epochs with batch size 512 on 8 GPUs.

For verifying the performance on different structures, we apply ODCS on the ResNet-34 with basic block and ResNet-50 with bottleneck block. As shown in Table 4, our ODCS gains substantial improvement based on a high baseline, 1.93%, 0.27% and 0.27% in ResNet-34, 50 and 101, respectively. To compare with related works, we list some results of SOTA methods, the ONI (Huang et al. 2020) and methods based on MHE (Liu et al. 2018), in Table 5. Significantly, the absolute value of our results is much higher than the others.

Method	34	50	101
Baseline(ResNet)	74.37	76.73	76.60
ODCS	76.30	77.00	76.87

Table 4: Accuracy of ImageNet-2012 dataset using ResNet-34, 50 and 101 with our ODCS.

Ablation Studies

We have also conducted ablation experiments on the structure based on our strategy and regularization terms on the CIFAR datasets.

Impact of nonlinear operation (ReLU). ReLU between two Convs have a significant effect according to our ODCS. Eliminating nonlinear operations is the key to make two Conv be linearly combined. We compare the performance of networks with and without ReLU between Convs. The results in Table 6 show that ReLU reduces models' performance for over 0.8%, in our ODCS. As for the reason, ReLU might prevent the linear combination between Convs to synthesize a feature extractor with better performance.

Method	34	50
MHE†	70.40	74.98
HS-MHE†	70.50	75.02
RP-CoMHE†	70.62	75.49
AP-CoMHE†	70.68	75.47
ONI*	-	76.45
ODCS	**76.30**	**77.00**

Table 5: Accuracy of ImageNet-2012 dataset with related methods. "†" means the data of (Liu et al. 2018). "*" means the data of (Huang et al. 2020).

Method	20	56	110
With ReLU	92.28	93.45	93.01
Without ReLU	**93.13**	**94.29**	**94.55**

Table 6: Accuracy comparison on CIFAR-10 with / without ReLU between Convs. The regularization L_{ODCS} is still working on the dictionary Conv in this experiment.

Weight for regularization. As shown in Table 7, we compare the influence of weight λ for loss function in ODCS. We test different λ settings from 5e-2 to 1e-3 on CIFAR with ResNet-20, 56 and 101. In Table 7, the standard deviations σ of the results of the models are all extremely small. The accuracy of the same network in the same dataset with different λ fluctuate slightly, indicating that nearly all networks are extremely insensitive to λ.

Specifically, we find the accuracy is slightly higher when λ is 5e-3 and we finally set the λ to 5e-3 for all experiments.

	CIFAR-10			CIFAR-100		
λ	20	56	110	20	56	110
5e-2	92.95	94.33	94.42	79.67	82.75	82.35
1e-2	92.81	94.10	94.37	79.43	82.48	82.70
5e-3	93.13	94.29	94.55	79.50	82.65	82.48
1e-3	92.88	94.30	94.32	79.48	82.54	82.36
σ	0.137	0.105	0.098	0.104	0.120	0.162

Table 7: Accuracy comparison on CIFAR-10 and CIFAR-100 with different weights for loss. All weights are tested on ResNet-20, 56 and 110. The standard deviation of accuracy σ is listed to show the stability for different weights.

Stable training without complex adjustment for loss weight. We carefully inspect the training curves of ResNet-110 on CIFAR-10 and ResNet-50 on CUB-200-2011 in Figure 3. Evidently, the weight of loss for ODCS is not sensitive. However, for other methods, the weight of loss need to be designed carefully to maintain. Besides, the schedule varies for different configurations, which makes the training more difficulty.

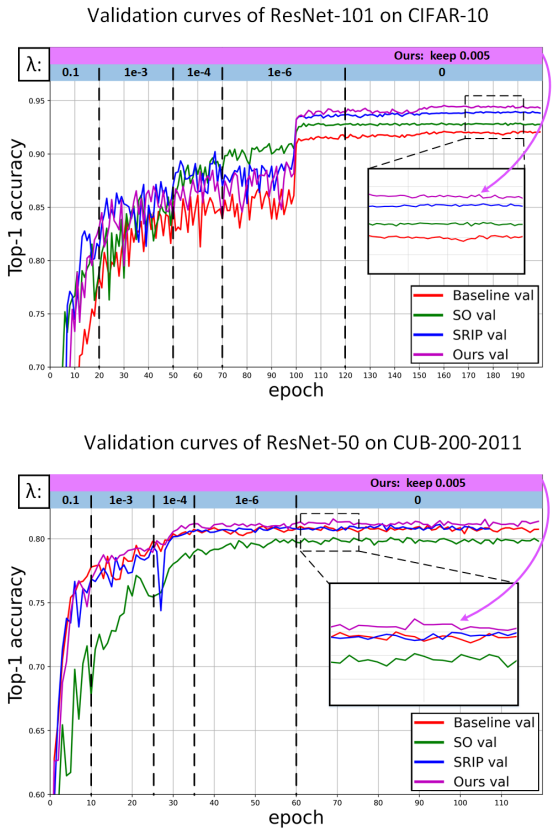

Figure 3: Training curves for different methods. Top: Training curves for ResNet-110 on CIFAR-10; Bottom: Training curves for ResNet-50 on CUB-200-2011. The tuning schedule of λ is shown on the top of the graph. We zoom in the optimal accuracy on curves to show the advantage of ODCS.

It is also observed that all the regularization methods significantly accelerate the training process in the initial stage, especially for ODCS and SRIP. After the second decrease of learning rate, ODCS still maintains the highest accuracy.

Influence of limiting norm in strategies. To solve the problem of being too restrictive in present regularization, we use loss L_{ODCS} instead of SO. To verify the validity of removing limit, we test the regularization with or without limit on the norm of vectors in W^TW by changing the corresponding loss functions.

The results in Table 8 show that ODCS is slightly worse than SO and SRIP when enforce norm=1. It's probably because ODCS only constrains kernels to be orthogonal and it is more relaxed than traditional regularization. In the early stage of training, the ODCS is not strict enough, which results in slow pre-training fitting with traditional strategy. The mismatch among the regular term, the object of action and the network structure causes its performance on the classification to decline. When the norm is free, the difference among ODCS, SO and SRIP is obvious, and our ODCS is slightly better than the others. In the strategy we proposed, orthogonality between kernels is a sufficient condition to extract enough features for dictionary Conv.

Compared to the same regularization in different strategy, our strategy is significantly better than traditional ones. SO increase more than 0.6%, and SRIP increase about 0.3%. Our ODCS achieves the largest margin, 1.67%, 1.79%, 2.05% on ResNet-20, 56 and 110 respectively. The experiment demonstrates that loss L_{ODCS}, as a basic conditions of orthogonality, is a suitable component in ODCS.

norm	Method	20	56	110
1	SO	92.46	93.59	93.46
	SRIP	92.65	93.89	93.96
	ODCS	91.46	92.50	92.45
free	SO	93.04	**94.29**	94.49
	SRIP	93.00	94.11	94.32
	ODCS	**93.13**	**94.29**	**94.55**

Table 8: Accuracy comparison on CIFAR-10 with different regularization loss functions. The influence of whether limits the norm of kernel to 1 is different for each strategy.

In summary. It's worth noted that the ablation studies show that our ODCS is an integrated method. Each step cannot be separated as an improvement method alone. Creating or locating the specific structure and removing ReLU is for Conv to form the pair of dictionary and encoding. Releasing the constraint on norm of kernels in regularization is to reduce the superfluous information in dictionary Conv kernels.

Conclusion and Future Works

Inspired by orthogonality regularization and sparse dictionary learning, an Orthogonal Dictionary Convolution Strategy (ODCS) is presented to improve the performance of CNNs. In this paper, we propose to changing the architectures by removing nonlinear layer between Convs and then imposing orthogonality regularization on specific dictionary Conv. As shown in experiments, ODCS could be easily applied to classical deep neural networks for various tasks. The extensive experiments show that our method achieves higher accuracy, more stable training curve and faster convergence than traditional strategy. Moreover, releasing the constraint on norm of kernels in L_{ODCS} not only enhances the diversity of dictionary Conv but also relaxes the too strict regularization in training. This makes the hyper-parameters tuning of regularization is more flexible. In the future work, we will explore more applications for ODCS, such as Image Denoising, 3D Point Cloud Detection.

Acknowledgements

This work was supported by National Key R&D Program of China (Grant No. 2018YF70202800), the Natural Science Foundation of China (NSFC) (Grant Nos. 61674120), JST, ACT-X (Grant No. JPMJAX190D), Japan and JST Moonshot R&D (Grant No. JPMJMS2011), Fundamental Research Funds for Central Universities and Innovation Fund of Xidian University.

References

Bansal, N.; Chen, X.; and Wang, Z. 2018. Can We Gain More from Orthogonality Regularizations in Training Deep Networks? In Bengio, S.; Wallach, H. M.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, 4266–4276.

Bengio, Y.; Simard, P. Y.; and Frasconi, P. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Networks*, 5(2): 157–166.

Brock, A.; Donahue, J.; and Simonyan, K. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Dauphin, Y. N.; Pascanu, R.; Gülçehre, Ç.; Cho, K.; Ganguli, S.; and Bengio, Y. 2014. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N. D.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, 2933–2941.

Glorot, X.; and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In Teh, Y. W.; and Titterington, D. M., eds., *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, volume 9 of *JMLR Proceedings*, 249–256. JMLR.org.

Harandi, M.; and Fernando, B. 2016. Generalized BackPropagation, Étude De Cas: Orthogonality. *CoRR*, abs/1611.05927.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 770–778. IEEE Computer Society.

Heide, F.; Heidrich, W.; and Wetzstein, G. 2015. Fast and flexible convolutional sparse coding. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 5135–5143. IEEE Computer Society.

Huang, G.; Liu, Z.; van der Maaten, L.; and Weinberger, K. Q. 2017. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2261–2269. IEEE Computer Society.

Huang, L.; Liu, L.; Zhu, F.; Wan, D.; Yuan, Z.; Li, B.; and Shao, L. 2020. Controllable Orthogonalization in Training DNNs. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 6428–6437. Computer Vision Foundation / IEEE.

Huang, L.; Liu, X.; Lang, B.; Yu, A. W.; Wang, Y.; and Li, B. 2018. Orthogonal Weight Normalization: Solution to Optimization Over Multiple Dependent Stiefel Manifolds in Deep Neural Networks. In McIlraith, S. A.; and Weinberger, K. Q., eds., *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, 3271–3278. AAAI Press.

Ioffe, S.; and Szegedy, C. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Bach, F. R.; and Blei, D. M., eds., *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, 448–456. JMLR.org.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In Bartlett, P. L.; Pereira, F. C. N.; Burges, C. J. C.; Bottou, L.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, 1106–1114.

Lee, H.; Battle, A.; Raina, R.; and Ng, A. Y. 2006. Efficient sparse coding algorithms. In Schölkopf, B.; Platt, J. C.; and Hofmann, T., eds., *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, 801–808. MIT Press.

Lin, R.; Liu, W.; Liu, Z.; Feng, C.; Yu, Z.; Rehg, J. M.; Xiong, L.; and Song, L. 2020. Regularizing Neural Networks via Minimizing Hyperspherical Energy. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 6916–6925. Computer Vision Foundation / IEEE.

Liu, C.; Wan, F.; Ke, W.; Xiao, Z.; Yao, Y.; Zhang, X.; and Ye, Q. 2019. Orthogonal Decomposition Network for Pixel-Wise Binary Classification. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 6064–6073. Computer Vision Foundation / IEEE.

Liu, W.; Lin, R.; Liu, Z.; Liu, L.; Yu, Z.; Dai, B.; and Song, L. 2018. Learning towards Minimum Hyperspherical Energy. In Bengio, S.; Wallach, H. M.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, 6225–6236.

Mailhé, B.; Lesage, S.; Gribonval, R.; Bimbot, F.; and Vandergheynst, P. 2008. Shift-invariant dictionary learning for sparse representations: Extending K-SVD. In *2008 16th European Signal Processing Conference, EUSIPCO 2008, Lausanne, Switzerland, August 25-29, 2008*, 1–5. IEEE.

Mairal, J.; Bach, F. R.; Ponce, J.; and Sapiro, G. 2009. Online dictionary learning for sparse coding. In Danyluk, A. P.; Bottou, L.; and Littman, M. L., eds., *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, volume 382 of *ACM International Conference Proceeding Series*, 689–696. ACM.

Mishkin, D.; and Matas, J. 2016. All you need is a good init. In Bengio, Y.; and LeCun, Y., eds., *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Miyato, T.; Kataoka, T.; Koyama, M.; and Yoshida, Y. 2018. Spectral Normalization for Generative Adversarial Networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Olshausen, B. A.; and Field, D. J. ???? Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*.

Ozay, M.; and Okatani, T. 2016. Optimization on Submanifolds of Convolution Kernels in CNNs. *CoRR*, abs/1610.07008.

Rabouy, C.; Paris, S.; and Glotin, H. 2015. Improving image classification by orthogonality of sparse codes. In Köppen, M.; Xue, B.; Takagi, H.; Abraham, A.; Muda, A. K.; and Ma, K., eds., *7th International Conference of Soft Computing and Pattern Recognition, SoCPaR 2015, Fukuoka, Japan, November 13-15, 2015*, 103–110. IEEE.

Ramírez, I.; Sprechmann, P.; and Sapiro, G. 2010. Classification and clustering via dictionary learning with structured incoherence and shared features. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, 3501–3508. IEEE Computer Society.

Rodríguez, P.; Gonzàlez, J.; Cucurull, G.; Gonfaus, J. M.; and Roca, F. X. 2017. Regularizing CNNs with Locally Constrained Decorrelations. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. S.; Berg, A. C.; and Li, F. 2015. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.*, 115(3): 211–252.

Sarhan, M. H.; Navab, N.; Eslami, A.; and Albarqouni, S. 2020. Fairness by Learning Orthogonal Disentangled Representations. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J., eds., *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIX*, volume 12374 of *Lecture Notes in Computer Science*, 746–761. Springer.

Saxe, A. M.; McClelland, J. L.; and Ganguli, S. 2014. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In Bengio, Y.; and LeCun, Y., eds., *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Tropp, J. A.; and Gilbert, A. C. 2007. Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit. *IEEE Trans. Inf. Theory*, 53(12): 4655–4666.

Wang, J.; Chen, Y.; Chakraborty, R.; and Yu, S. X. 2020. Orthogonal Convolutional Neural Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 11502–11512. IEEE.

Xie, D.; Xiong, J.; and Pu, S. 2017. All You Need is Beyond a Good Init: Exploring Better Solution for Training Extremely Deep Convolutional Neural Networks with Orthonormality and Modulation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 5075–5084. IEEE Computer Society.

Xie, S.; Girshick, R. B.; Dollár, P.; Tu, Z.; and He, K. 2017. Aggregated Residual Transformations for Deep Neural Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 5987–5995. IEEE Computer Society.

Yaghoobi, M.; Blumensath, T.; and Davies, M. E. 2008. Regularized dictionary learning for sparse approximation. In *2008 16th European Signal Processing Conference, EUSIPCO 2008, Lausanne, Switzerland, August 25-29, 2008*, 1–5. IEEE.

Yanai, K.; Tanno, R.; and Okamoto, K. 2016. Efficient Mobile Implementation of A CNN-based Object Recognition System. In Hanjalic, A.; Snoek, C.; Worring, M.; Bulterman, D. C. A.; Huet, B.; Kelliher, A.; Kompatsiaris, Y.; and Li, J., eds., *Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016, Amsterdam, The Netherlands, October 15-19, 2016*, 362–366. ACM.

Yang, H.; Tang, M.; Wen, W.; Yan, F.; Hu, D.; Li, A.; Li, H.; and Chen, Y. 2020. Learning Low-rank Deep Neural Networks via Singular Vector Orthogonality Regularization and Singular Value Sparsification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*, 2899–2908. IEEE.

Zagoruyko, S.; and Komodakis, N. 2016. Wide Residual Networks. In Wilson, R. C.; Hancock, E. R.; and Smith, W. A. P., eds., *Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016*. BMVA Press.

Zhou, J.; Do, M. N.; and Kovacevic, J. 2006. Special paraunitary matrices, Cayley transform, and multidimensional orthogonal filter banks. *IEEE Trans. Image Process.*, 15(2): 511–519.