V_kD : Improving Knowledge Distillation using Orthogonal Projections

Roy Miles Ismail Elezi Jiankang Deng
Huawei Noah's Ark Lab

2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3503-5300-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/CVPR52733.2024.01488

Abstract

Knowledge distillation is an effective method for training small and efficient deep learning models. However, the efficacy of a single method can degenerate when transferring to other tasks, modalities, or even other architectures. To address this limitation, we propose a novel constrained feature distillation method. This method is derived from a small set of core principles, which results in two emerging components: an orthogonal projection and a task-specific normalisation. Equipped with both of these components, our transformer models can outperform all previous methods on ImageNet and reach up to a 4.4% relative improvement over the previous state-of-the-art methods. To further demonstrate the generality of our method, we apply it to object detection and image generation, whereby we obtain consistent and substantial performance improvements over state-of-the-art. Code and models are publicly available[1].

1. Introduction

Deep learning has achieved remarkable success across a wide variety of tasks in computer vision [38], audio [58], and language [17] domains. However, its adoption is often coupled with increasing computational costs which has limited its application on resource constrained devices such as mobile phones. Fortunately, there have been many techniques proposed to train fast and efficient networks, such as weight pruning [24, 39, 41], quantisation [6, 88], and knowledge distillation [9, 34, 47, 69].

Knowledge distillation (KD) in particular has shown great success [4, 20, 45, 50, 69]. Its main idea is to utilise the pre-trained knowledge of a much larger (teacher) model to supervise the training of a much smaller (student) model. Traditional KD [3, 34] methods have focused on image classification by using the softmax predictions of the teacher as ground-truth labels for the student. However, doing so has made them limited and only applicable to specific modalities, or tasks. Although feature distillation [8, 69] can relax this constraint on the downstream task or modality, its adoption can incur significant computational costs due to the construction of expensive relational objects [49, 50, 53] and memory banks [69]. Most feature distillation pipelines can be described as using some projection [64], alignment [12], or fusion module [9]. These components, though widely used, tend to rely on heuristic design choices. Such heuristic-driven approaches often fall short in delivering new insights into the underlying mechanics of distillation and struggle to adapt to diverse tasks without the introduction of several auxiliary losses. [9, 16, 44, 85]. Furthermore, these additional losses introduce additional hyperparameters which will require tuning for specific tasks or settings.

In this work, we propose a novel projection layer that is derived from a principal concept. Our approach focuses on one key idea: Preserving the intra-batch feature similarity. We highlight that if the similarity between features is preserved, then the projection layer will not change or alter the underlying student representation. This constraint is important since it will maximise the amount of knowledge being distilled to the student backbone. For example, if the projector is too expressive it may shortcut the distillation objective by learning some complex non-linear mapping between the

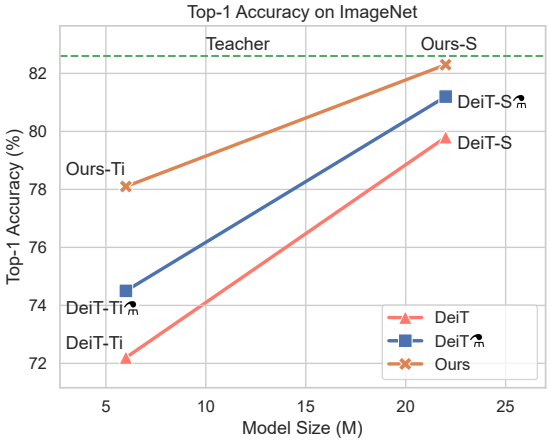


Figure 1. Comparison to both DeiT and DeiT⚗ [71] on ImageNet-1K, where DeiT⚗ is a distilled DeiT model using a distillation token. Our proposed distillation method achieves significant improvements over DeiT-Ti, while effectively bridging the gap between the teacher and student performance for DeiT-S.

[1] https://github.com/roymiles/vkd

2575-7075/24/$31.00 ©2024 IEEE
DOI 10.1109/CVPR52733.2024.01488

two spaces. This result would significantly diminish the efficacy of distillation and is especially detrimental since the projector is thrown away after training.

By enforcing the preservation of the feature similarity, we derive a reparameterisation of the projection layer itself using the set of orthogonal matrices. We propose to efficiently implement this reparameterisation by projecting the weights onto $SO(n)$ and then truncating the excess rows. In doing so, we enforce the property of row-wise orthogonality, while avoiding the need to compute any expensive matrix inversions or factorisations. We show that this constraint not only improves the student performance but also improves the training convergence and the efficacy of distilling inductive biases.

Finally, a common component in the application of knowledge distillation for generative tasks is the use of additional auxiliary losses. These losses can encourage the generation of diverse features, which will subsequently lead to the generation of more diverse images. However, these auxiliary losses often conflict with the distillation objective and will subsequently degrade the student performance. To address this limitation, we propose a unified framework for incorporating these auxiliary objectives into the distillation loss itself using a task-specific normalisation step. Thus, we show that simply whitening the teacher features can implicitly encourage feature diversity, while removing the need for fine-tuning the hyperparameters of many additional losses. We further demonstrate the importance of this whitening step for data-limited image generation, whereby we achieve a consistent and substantial performance improvements over state-of-the-art. In summary, our contributions are outlined as follows:

- We propose a novel orthogonal projection layer to maximise the knowledge being distilled through to the student backbone.
- We complement our projection with a task-wise normalisation that enables knowledge distillation in generative tasks.
- We apply our method to a wide range of tasks and modalities, improving over the state-of-the-art by up to 4.4% on ImageNet-1K (see Fig. 1).

2. Related work

Knowledge Distillation utilises the knowledge of a pretrained model as supervision for a much smaller model, which can then enable the application and deployment within resource constrained environments. The field can be broadly divided into two main areas: logits distillation [3, 14, 34, 37, 52] and feature distillation [9, 31, 49, 64, 69, 79, 80]. Logit distillation focuses on classification based tasks and introduces an additional objective to minimise the distance between the student and teacher predictions. This was originally proposed using the KL divergence [34], however it has since been extended using spherical normalisation [23], label decoupling [86], and probability reweighting [52]. In our work, we focus on feature distillation due to its generality to other tasks [9, 50] and modalities [65, 81]. Unfortunately, there is no underlying metric for the intermediate representation spaces, which has led to many heuristically derived solutions. For example, the hand-crafted FSP matrices were proposed [83] to capture the relation between features before and after a set of residual layers. Similarly, many other works have proposed to transfer knowledge using the construction of various Gram [28, 43, 74] or correlation-based matrices [40, 49, 57]. The activation boundaries have also been shown to be an effective supervisory signal for distillation [32], along with the gradients to capture the loss landscape [68, 89]. Another line of work can be loosely grouped together by their inspirations from the self-supervision [69, 77] or information theory literature [49, 50]. In contrast to these methods, we take a step back from the conventional curation of hand-crafted relational objects or objectives. We instead derive and demonstrate that an orthogonal projection is much more effective and can be used in conjunction with just a simple $L2$ loss.

Self-supervised learning describes the family of pretext tasks used to learn good representations of data in the absence of any ground truth labels. This is an important topic with the overwhelming abundance of unlabelled data available and the increasing costs for human annotations. Self-supervised learning shares some significant similarities with the knowledge distillation literature. For example, contrastive learning [10, 29, 35] has already inspired many distillation methods [8, 69, 78] and asymmetric architectures [11, 21, 22, 62] are often described as a form of self-distillation [2]. However, the most salient overlap with our work instead lies with the use of a predictor. The predictor is a learnable model that maps from the online network to the momentum network. Its usage is very similar to what is described as a projector [13] in the knowledge distillation literature. DirectPred [70] explored the training dynamics of this predictor, which allowed the derivation of a closed form solution for its weights. This work was later simplified by either removing the expensive eigen-decomposition [76] or using fast matrix iterations [63]. In contrast to these works, we explore the role of the projector for knowledge distillation. By building upon a simple set of principles for KD, we are able to derive a cheap reparameterisation of the projector weights that can maximise the knowledge transfer.

Layer reparameterisation has been widely adopted as a technique for constraining weights to introduce favourable properties. For instance, unitary matrices have been shown to address the gradient issues in RNNs [1], positive definite matrices enhance the robustness of batch normalization layers [5], and orthogonal matrices offer spectral regularization for improved generalization [1, 30]. More recent works

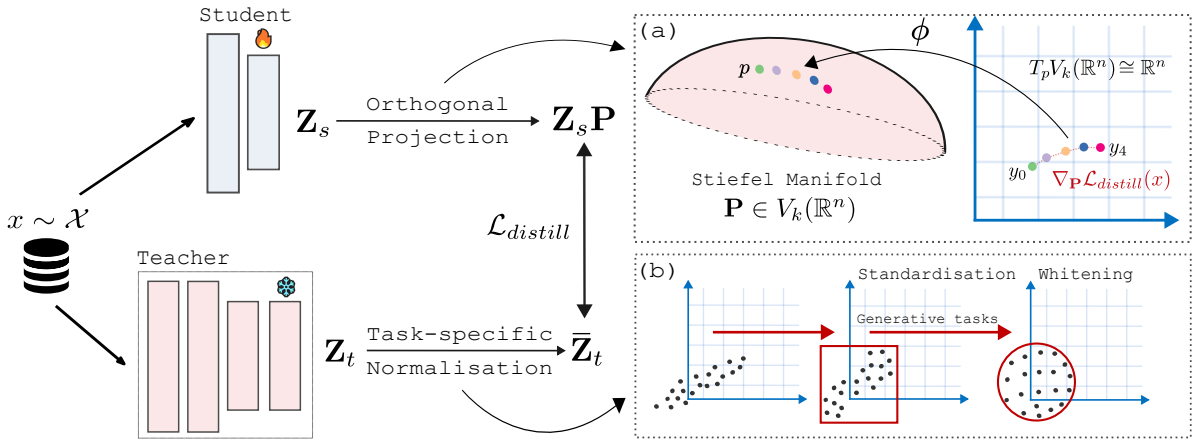


Figure 2. Illustration of our proposed feature distillation using an orthonormal projection and task-specific feature normalisation. The orthonormal projection (a) maximises the knowledge being distilled to the student backbone, while the task-specific normalisation (b) can introduce domain-specific priors to improve model performance. 🔥 denotes trainable weights, while ❄ denotes weights which are frozen.

have shown that low-rank matrices are effective in reducing the cost of fine-tuning large language models [36], while orthogonal matrices enable the cheap controllable fine-tuning of text-image diffusion models [59]. In our work we take these ideas into the context of knowledge distillation with an orthogonal projection. We show that this orthogonal constraint improves both the efficacy of distillation and improves the overall model convergence.

3. Orthogonal Projections

Despite the generality of feature distillation, its use is often coupled with many design decisions and heuristics. These decisions arise from the construction of multiple losses between intermediate feature maps which incur significant and unnecessary training overheads. To address these constraints, we adopt a simple feature distillation pipeline (see Fig. 2) using only the features directly before the classifier, or in the case of generative tasks, the latent representation.

In section 3.1 we motivate the necessary conditions to maximise the efficacy of distillation through the projection. This leads to a reparameterisation of the projection as an orthogonal matrix, which is then efficiently implemented in section 3.2. In section 3.3 we provide some interesting additional insights into the properties of these orthogonal projections, while in section 3.4 we extend our distillation pipeline to improve the performance on both generative and discriminative tasks by using an additional task-specific normalisation step.

3.1. Why use orthogonal projections?

Our main objective is to mitigate the possibility of the projection layer learning any new representation of the data that is not shared by the feature extractor. This is important because the projection layer is thrown away after training and we want to match the feature extractor with the teacher, rather than solely matching the projected features. To achieve this, we propose to preserve the structural information through the projection. We describe this structural information using a kernel matrix $\mathbf{K} \in \mathbb{R}^{b\times b}$, where b is the batch-size. This kernel matrix captures the pairwise similarity between all features within a batch:

$$\mathbf{K}_{ij} = k(\mathbf{Z}^s_i, \mathbf{Z}^s_j) = \langle \mathbf{Z}^s_i, \mathbf{Z}^s_j \rangle_{\mathcal{H}}\,, \tag{1}$$

where $\mathcal{H}$ is some Hilbert space implicitly defined by the positive-definite real-valued kernel k and $\mathbf{Z}^s \in \mathbb{R}^{b\times d_s}$ are the student features of dimension d_s. We aim to preserve $\mathbf{K}$ under the application of a linear transformation of its arguments. This is equivalent to preserving the structural information of the features. We can express many practical kernels, such as the radial basis function kernel or the polynomial kernel, using a Taylor series expansion [15]:

$$k(\mathbf{Z}^s_i, \mathbf{Z}^s_j) = \sum_{n=0}^{\infty} a_n \langle \mathbf{Z}^s_i, \mathbf{Z}^s_j \rangle^n\,, \tag{2}$$

where a_n are the coefficients. This expression shows that we simply need a transformation $\mathbf{P}$ that preserves inner products. Using the canonical inner product in $\mathbb{R}^{d_s}$, we derive this constraint on $\mathbf{P}$ as follows:

$$\mathbf{Z}^s_i(\mathbf{Z}^s_j)^T = \mathbf{Z}^s_i\mathbf{P}(\mathbf{Z}^s_j\mathbf{P})^T \tag{3}$$

$$= \mathbf{Z}^s_i\mathbf{P}\mathbf{P}^T(\mathbf{Z}^s_j)^T, \tag{4}$$

which holds if $\boxed{\mathbf{P}^T = \mathbf{P}^{-1}}$. This constraint conveniently defines the special orthogonal group $SO(d_s)$ with $d_s = d_t$. This group parameterises the set of all rotations in $\mathbb{R}^{d_s}$, thus very naturally and intuitively preserves the idea of structural information. However, in the general case where $d_s \neq d_t$, our projection matrix $\mathbf{P}$ is no longer square. This means

that there exists no canonical inverse for our derived constraint to hold. Choosing the right-inverse defines the set of matrices with orthonormal rows, whereas choosing the left-inverse defines the set of matrices with orthonormal columns. Motivated by the need for an efficient reparameterisation, we focus our attention on the right-inverse, which defines the set of matrices with orthonormal rows, its transpose of which is conveniently represented as a Stiefel matrix manifold [18], denoted $V_{d_t}(\mathbb{R}^{d_s})$. To simplify further notation, we omit this distinction between the two.

Since the Stiefel manifold is smooth, it facilitates the use of standard gradient descent techniques using reparameterisations[2]. In the next section we provide an efficient implementation of this reparameterisation.

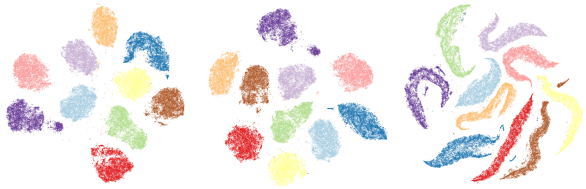
3.2. Orthogonal reparameterisation

There are a few convenient ways to ensure orthogonality of the projection matrix $\mathbf{P}$. One of these ways is to use a Cayley transformation [25] that constructs an orthogonal matrix $\mathbf{P}$ from a skew-symmetric matrix: $\mathbf{P} = (\mathbf{I}-\mathbf{W})(\mathbf{I}+\mathbf{W})^{-1}$ where $\mathbf{W} = -\mathbf{W}^T$. Unfortunately, this closed-form parameterisation, despite its simplicity, requires the expensive computation of a large matrix inverse. Using a QR decomposition can also be considered, but will require the use of expensive iterative algorithms, such as the Gram Schmidt process. Since an orthogonal reparameterisations map is needed for each iteration during training, it is critical for its evaluation to be computationally cheap. To address this computational constraint we propose an efficient algorithm that avoids the need for any expensive matrix inversions or factorisations. We propose to instead perform a cheap parameterisation map onto $SO(d_t)$ using the matrix exponential $exp(\mathbf{A})$, which can be efficiently implemented using the padé approximation. Knowing that $\mathbf{W}$ is skew-symmetric, we show that $\exp(\mathbf{W})$ is an orthogonal matrix using a few properties of the exponential: $\exp(\mathbf{W}) \cdot \exp(\mathbf{W})^T = \exp(\mathbf{W}+\mathbf{W}^T) = \exp(-\mathbf{W}^T+\mathbf{W}^T) = \exp(\mathbf{0}) = \mathbf{I}$. We then project back to $V_{d_t}(\mathbb{R}^{d_s})$ by dropping the last $d_t - d_s$ rows. These two sequential steps are given more compactly as follows::

$$\phi : \mathbf{W} \xrightarrow{\exp(\mathbf{W})} \mathbf{A} \in SO(d_t) \xrightarrow{\mathbf{A}_{:d_s}} \mathbf{P} \in V_{d_t}(\mathbb{R}^{d_s}), \quad (5)$$

where the subscript notation follows from the slicing convention in Pytorch [56] and Numpy [27]. This detour only requires the computation of one exponential, which can be cheaply evaluated using the Padé approximation [33]. We show an illustration of this re-parameterisation in Fig. 2a with the map ϕ described in equation 5.

[2]Reparameterisations can be seen as surjective functions that map from Euclidean space back onto the manifold.



(a) original (b) orthogonal (c) linear

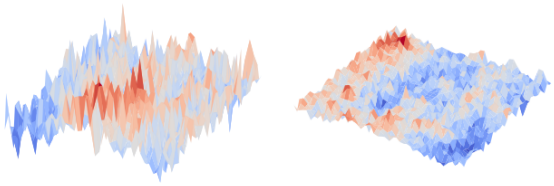
Figure 3. t-SNE visualisation [75] of features undergoing either a linear or orthogonal transformation. The orthogonal transformation preserves all of the structural feature information, whereas the linear projection can distort a lot of structure, which can diminish the efficacy of distillation.

3.3. Orthogonal projections minimise redundancy

The orthogonal transformations ensure that the projected features $\mathbf{Z}^s\mathbf{P}$ will not only be a linear combination of the original features, but also be a transformation that preserves the notion of distance between features. This result can be explained more concretely by observing that the singular values of $\mathbf{P}$ are all 1. Geometrically, this means that the projection is not squashing or distorting along any of the dimensions, which would bias the loss toward reconstructing the teachers features using only a subset of features. We illustrate this phenomenon in Fig. 3 where we find that even a simple linear projection distorts the features making them overlap with each other. This is caused because it attempts to align the space with the teacher. This level of distortion can degrade the linear separability of the features for the classifier, which impacts the model performance. In contrast, the orthogonal projection preserves the underlying feature manifold.

3.4. Introducing domain-specific priors

For many tasks it is important to invoke domain specific priors or auxiliary losses to improve model performance [16, 44]. Unfortunately, many of these auxiliary losses conflict with the distillation objective and hinder its efficacy. Instead, we propose a general framework for normalisation that naturally and implicitly incorporates these priors into the distillation objective itself. We show that standardisation is very effective for discriminative tasks by improving the model convergence. This convergence property can be attributed to the improved robustness of the distillation loss to random input perturbations. Similarly, we also show that whitening is a critical step for generative tasks by providing an implicit and soft encouragement of diverse features, which has been proven effective for generating diverse images [19]. Whitening can be much more effective and significantly cheaper than introducing additional auxiliary losses, which has been previously proposed in the literature [16]. We now provide a more detailed illustration of these ideas.



77.7% top-1 Acc
(a) without normalisation

78.3% top-1 Acc
(b) **with** normalisation

Figure 4. Visualisation of the V_kD-Ti $\mathcal{L}_{distill}$ loss landscape with perturbations of the input image across two random dimensions. Normalisation significantly reduces the sensitivity of the loss to random perturbations, which leads to improved robustness and convergence for training.

Standardisation improves model convergence. In our application of knowledge distillation to discriminative tasks, we observe that a straightforward normalisation of the teacher's representation yields a notable improvement in the robustness of the distillation loss to spurious deformations of the input image (see Fig. 4). These spurious deformations arise from the increasingly expressive families of data augmentation strategies being commonly employed for knowledge distillation [71]. We find that minimizing this loss variance can significantly improve the overall model convergence and performance.

Whitening improves feature diversity. By whitening the teacher features, we derive a lower bound that resembles a feature diversity loss [16]. We start with an $L2$ loss between $\mathbf{Z}_s\mathbf{P} \in \mathbb{R}^{b\times d}$ and $\mathbf{Z}_t \in \mathbb{R}^{b\times d}$. Since $\mathbf{P}$ is an orthogonal projection that preserves inner products (see section 3.1), we omit its usage to simplify analysis:

$$\begin{aligned}\mathcal{L}_{distill} &= \left\|\mathbf{Z}^s - \mathbf{Z}^t\right\|^2 \\ &= \sum_{i\neq j}\left\|\mathbf{Z}^s_{:,j} - \mathbf{Z}^t_{:,i} - \mathbf{Z}^t_{:,j} + \mathbf{Z}^t_{:,i}\right\|^2 \\ &= \sum_{i\neq j}\left\|\mathbf{Z}^s_{:,j} - \mathbf{Z}^t_{:,i}\right\|^2 + \left\|\mathbf{Z}^t_{:,j} + \mathbf{Z}^t_{:,i}\right\|^2 \\ &\quad - 2\left\langle \mathbf{Z}^s_{:,j} - \mathbf{Z}^t_{:,i}, \mathbf{Z}^t_{:,j} + \mathbf{Z}^t_{:,i}\right\rangle \end{aligned} \tag{6}$$

Since $\mathbf{Z}^t$ is whitened such that $(\mathbf{Z}^t)^T(\mathbf{Z}^t) = \mathbf{I}$, i.e., perfect decorrelation of features, we can significantly simplify the expression above:

$$\begin{aligned}\mathcal{L}_{distill} &= \sum_{i\neq j}\left\|\mathbf{Z}^s_{:,j} - \mathbf{Z}^t_{:,i}\right\|^2 + 2 - 2\left\langle \mathbf{Z}^s_{:,j} - \mathbf{Z}^t_{:,i}, \mathbf{Z}^t_{:,j} + \mathbf{Z}^t_{:,i}\right\rangle \\ &\geq \sum_{i\neq j}\left\|\mathbf{Z}^s_{:,j} - \mathbf{Z}^t_{:,i}\right\|^2 + 2 - 2\left\|\mathbf{Z}^s_{:,j} - \mathbf{Z}^t_{:,i}\right\|^2\left\|\mathbf{Z}^t_{:,j} + \mathbf{Z}^t_{:,i}\right\|^2 \\ &= \boxed{\text{const} - \lambda\sum_{i\neq j}\mathbf{C}^2_{j,i}}, \quad \text{where} \quad \mathbf{C}_{i,j} = \left\|\mathbf{Z}^s_{:,j} - \mathbf{Z}^t_{:,i}\right\| \end{aligned} \tag{7}$$

where *const* and $\lambda \geq 3$ are both constants that do not depend on the model parameters. Here $\mathbf{C}$ is the euclidean cross-correlation matrix that captures the distance between all the pairs of student and teacher features. This derivation simply shows that minimising an $L2$ loss subject to an explicit whitening constraint on the teacher features provides a cross-feature objective. This cross-feature objective maximises the off-diagonal entries in the cross-correlation matrix, thus encouraging all the features to be decorrelated with respect to the teacher. We describe this process of encouraging decorrelation of features as increasing the feature diversity. We validate this connection of whitening with feature diversity in section 4.3, where we employ knowledge distillation in the context of data-efficient image generation.

4. Experiments

In this section, we evaluate the generality of our simple knowledge distillation pipeline across three distinct vision tasks: Image classification, object detection, and image generation. In each of these tasks, we consider the harder distillation settings, such as distilling cross-architecture, or in the data-efficient regimes. Throughout we use V_kD to denote our KD method using an orthogonal projection, i.e., a matrix projection from the Stiefel manifold $\mathbf{V_k}(\mathbb{R}^\mathbf{d})$.

Network	acc@1	Teacher	#params
RegNetY-160 [60] *CVPR20*	82.6	none	84M
CaiT-S24 [72] *ICCV21*	83.4	none	47M
DeiT3-B [73] *ECCV22*	**83.8**	none	87M
DeiT-Ti [71] *ICML21*	72.2	none	5M
CivT-Ti [61] *CVPR22*	74.9	regnety-600m + rednet-26	6M
MaskedKD [66] *arXiv23*	75.4	cait-s24	6M
Manifold [26] *NeurIPS22*	76.5	cait-s24	6M
DeiT-Ti⚗ [71] *ICML21*	74.5	regnety-160	6M
↳ 1000 epochs	76.6	regnety-160	6M
DearKD [12] *CVPR22*	74.8	regnety-160	6M
↳ 1000 epochs	77.0	regnety-160	6M
USKD [82] *ICCV23*	75.0	regnety-160	6M
SRD [48] *AAAI24*	77.2	regnety-160	6M
V_kD-Ti	**78.3**	regnety-160	6M
DeiT-S [71] *ICML21*	79.8	none	22M
CivT-S [61] *CVPR22*	82.0	regnety-4gf + rednet-50	22M
MaskedKD [66] *arXiv23*	81.4	deit3-b	22M
DeiT-S⚗ [71] *ICML21*	81.2	regnety-160	22M
↳ 1000 epochs	82.6	regnety-160	22M
DearKD [12] *CVPR22*	81.5	regnety-160	22M
↳ 1000 epochs	82.8	regnety-160	22M
USKD [82] *ICCV23*	80.8	regnety-160	22M
SRD [48] *AAAI24*	82.1	regnety-160	22M
V_kD-S	**82.3**	regnety-160	22M

Table 1. Data-efficient training of transformers using knowledge distillation on the ImageNet-1K dataset. Unless specified, each model is only trained for 300 epochs.

Implementation details. We train all models in Pytorch [55] using 2 NVIDIA V100 GPUs. For the ImageNet experiments, we follow DeiT [71] using the same training schedule and optimization parameters. We also adopt Mixup [84] augmentation, but replace rand-augment

Method	Backbone	Epochs	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	Param.	FPS
ViDT	Swin-nano	50	40.4	59.6	43.3	23.2	42.5	55.8	16M	20.0 (45.8)
	↳ w/ V_kD	50	**43.0**	**62.3**	**46.2**	**24.8**	**45.3**	**60.1**		
	Swin-tiny	50	44.8	64.5	48.7	25.9	47.6	62.1	38M	17.2 (26.5)
	↳ w/ V_kD	50	**46.9**	**66.6**	**50.9**	**27.8**	**49.8**	**64.6**		
	Swin-small	50	47.5	67.7	51.4	29.2	50.7	64.8	61M	12.1 (16.5)
	↳ w/ V_kD	50	**48.5**	**68.4**	**52.4**	**30.8**	**52.2**	**66.0**		

Table 2. Comparison with other detectors on COCO2017 val set. FPS is measured with batch size 1 of 800 × 1333 resolution on a single Tesla V100 GPU, where the value inside the parentheses is measured with batch size 4 of the same resolution to maximize GPU utilisation. All of the student models are distilled from a pre-trained ViDT-base.

Student	ViDT (Swin-nano)		ViDT (Swin-tiny)	
Teacher	ViDT (small)	ViDT (base)	ViDT (small)	ViDT (base)
No Distillation [67] *ICLR22*	40.4		44.8	
Token Matching [67] *ICLR22*	41.5	41.9	45.8	46.5
V_kD	**42.2**	**43.0**	**45.9**	**46.9**

Table 3. Comparison of ViDT on COCO2017 val set. We report AP for the student models distilled from different teacher models.

with random gray scaling, gaussian blurring, and solarization. We use AdamW [46] optimizer with learning rate set to 0.001 and weight decay to 0.05. For the object detection experiments, we follow the same training methodology as ViDT [67] except that we replace the original token matching loss with our V_kD. Finally, for the image generation task, we use the same training methodology as KD-DLGAN [16] except that we remove the auxiliary diversity losses and instead replace it with either teacher standardisation or whitening. We also remove any distillation from the text encoders, thus further reducing the cost of our method in comparison.

4.1. Data efficient training of transformers

We experiment with vision transformers, due to their proven success across a variety of fields. However, despite this success, they demand excessive training data and long training schedules. This limitation has motivated the use of knowledge distillation for improving the data efficiency of transformer models [71]. We compare our method with several others using the common knowledge distillation setting proposed alongside DeiT [71] that uses a CNN teacher pre-trained on ImageNet-21K. We train each student model for 300 epochs on ImageNet-1K for the image classification task. Unlike other methods that propose to leverage the efficacy of distilling through distillation tokens alone, we propose to distill directly through to the patch tokens. We present the results in Tab. 1 where we massively outperform the previous state-of-the-art. In the tiny architecture, we outperform the baseline by 6.1 percentage points (*pp*), and the previous best method that uses the same teacher, the USKD by 3.3*pp*, or a relative improvement of 4.4%. Interestingly, we perform better than DearKD trained for 1000 epochs by 1.3*pp*, showing that for distillation, it is not necessary to train that long. We reach similar results in the small architecture too, where we outperform all the other methods that use similar training resources, and reach competitive results with the methods that use more than 3 times as long training time. In fact, unlike other methods, our approach bridges the gap between the teacher model that reaches 82.6% accuracy without needing to introduce any excessively long training schedules.

4.2. Object detection

We consider the object detection task using the common MS-COCO benchmark [42]. We use the ViDT transformer architecture [67] due to its task performance and its efficiency on consumer hardware. We present the results in Tab. 2, and observe a significant and consistent improvement across a wide range of different ViDT variants. We improve using Swin-nano backbone by 2.6*pp*, in Swin-tiny backbone by 2.1*pp* and in Swin-small backbone by 1*pp*. Furthermore, we also compare against an alternative distillation method, described as token matching [67]. We present these results in Tab. 3. Our method outperforms token matching by up to 1.1*pp*, reaching the best results when we use a larger teacher, demonstrating that our method is not limited in the cases of larger capacity gaps.

4.3. Data limited image generation

To demonstrate the generality of our feature distillation framework, we consider an image generation task and compare to the recent KD-DLGAN [16]. KD-DLGAN proposes to use both the text and feature embeddings for the feature guidance followed by an additional diversity loss. Using our novel framework, we show that neither of these explicit additional losses is necessary. Instead, we use a simple whitening of features to encourage the generation of diverse images. We show these results in Tab. 4. A noticeable observation in these results is that whitening can obtain the most significant improvements in the more extreme data-limited regimes. For example, when training on 10% data with CIFAR-100, whitening the teacher features can improve the FID by up to 9.09. This result highlights that feature diversity is much more critical when there is insufficient training data. To show that our method is

Method	CIFAR-10			CIFAR-100		
	10% Data	20% Data	100% Data	10% Data	20% Data	100% Data
DA + KD (CLIP)	22.03 ± 0.07	13.70 ± 0.08	8.70 ± 0.02	33.93 ± 0.09	21.76 ± 0.06	11.74 ± 0.02
DA (Baseline)	23.34 ± 0.09	14.53 ± 0.10	8.75 ± 0.03	35.39 ± 0.08	22.55 ± 0.06	11.99 ± 0.02
KD-DLGAN	14.20 ± 0.06	11.01 ± 0.07	8.42 ± 0.01	18.03 ± 0.11	15.60 ± 0.08	$\mathbf{10.28 \pm 0.03}$
DA + V_kD	16.42 ± 0.07	10.94 ± 0.07	8.28 ± 0.01	24.92 ± 0.15	17.97 ± 0.17	10.61 ± 0.08
↳ w/ whitening	$\mathbf{13.16 \pm 0.06}$	$\mathbf{10.24 \pm 0.06}$	$\mathbf{8.20 \pm 0.03}$	$\mathbf{16.87 \pm 0.09}$	$\mathbf{14.00 \pm 0.07}$	10.41 ± 0.01

Table 4. Comparison with the state-of-the-art over CIFAR-10 and CIFAR 100. Competitive performance is achieved using the orthogonal projection alone, however, introducing a simple whitening step is sufficient in outperforming state-of-the-art by a significant margin. All the compared methods employ BigGAN as the backbone. FID is averaged over three runs.

Method	CIFAR-10 10% Data	CIFAR-100 10% Data
DA (Baseline) [87] *NeurIPS20*	23.34 ± 0.09	35.39 ± 0.08
FitNets [64] *ICLR14*	22.03 ± 0.07	33.93 ± 0.09
Label Distillation [34] *arXiv15*	20.46 ± 0.10	34.14 ± 0.11
PKD [54] *ECCV19*	21.34 ± 0.08	32.15 ± 0.13
SPKD [74] *ICCV19*	19.11 ± 0.07	31.97 ± 0.10
KD-DLGAN [16] *CVPR23*	14.20 ± 0.06	18.03 ± 0.11
V_kD	16.47 ± 0.07	24.92 ± 0.15
↳ w/ whitening	$\mathbf{13.16 \pm 0.06}$	$\mathbf{16.87 \pm 0.09}$

Table 5. Comparison to other knowledge distillation methods for image generation. Results were originally reported in [16] and highlight the importance of incorporating domain-specific priors - in this case, encouraging diverse features.

much more general than other KD methods in the literature, we also include a comparison for the hardest data-efficient regime in Tab. 5. The results show a consistent improvement in performance, which highlight the importance of engineering both the projector architecture and the normalisation scheme, as opposed to focusing on the distance metrics alone [16, 74].

4.4. Ablation study

We do a series of ablation studies to highlight the importance of our proposed building blocks. We also provide qualitative results that provide additional insights into explaining the efficacy of our distillation framework.

Effectiveness of orthogonal projections. To demonstrate the effectiveness of constraining the projection weights to be orthogonal, we consider the use of various other projection variants. We analyze the use of a projector ensemble [13], a multi-layer perceptron [51], and a standard linear layer. We present the results in Fig. 5. We observe that both the MLP and projector ensembles show improved performance over a linear layer when under short training schedules. However, when we extend the training schedule, the linear layer becomes much more effective. This is a consequence of the expressive projections beginning to learn new representations that are no longer shared by the student feature extractor. In contrast, our orthogonal projection not only improves the final accuracy but also improves the convergence properties for training, reaching state-of-the-art results in only ~ 200 epochs.

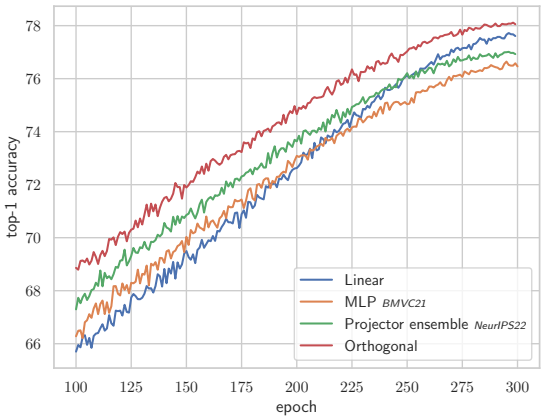


Figure 5. Comparing the performance and convergence of various projector reparameterisations. Although the MLP layer initially trains fast, it begins to saturate as it starts to learn a new representation of the data.

The effect of each block. We now disentangle the contribution of each block of our framework. In table 6 we report the final accuracy with and without normalisation or an orthogonal projection. We observe that, in the classification task, most of the performance improvement in obtained by our orthogonal projection. For example, the orthogonal projection alone boosts the performance from 76.3% to 77.9%. This observation is in contrast to the generative tasks, whereby we observe a necessity to use normalisation for strong performance.

Orth.	Norm.	Acc@1	Acc@5
✗	✗	76.3	93.3
✗	✓	76.9	93.5
✓	✗	77.9	93.9
✓	✓	**78.3**	**94.1**

Table 6. Highlighting the primary importance of an orthonormal projection. Image classification on ImageNet-1K using a DeiT-Ti student and a RegNety-160 teacher.

Whitening for generative tasks. To empirically confirm the importance of feature whitening for generative tasks, we perform an evaluation with and with it being used. We

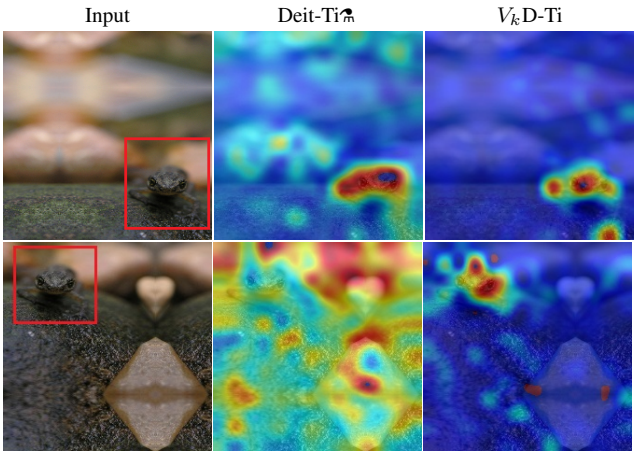


Figure 6. Evaluating the translational equivariance of attention maps. We select the best channel for the first translation and observe its attention maps after translating the input image again.

present the results in Tab. 4 and 5. We observe that not only is whitening necessary to outperform previous state-of-the-art image generation, but it also leads to a larger increase in performance in the data-limited regime. This result highlights the more prominent importance of diverse features when limited training data is available.

Distilling inductive biases. Knowledge distillation has proven effective for improving the data efficiency in training transformer models, especially when the teacher is a CNN. Unfortunately, there has been little qualitative analysis on explaining why this cross-architecture setting helps. We now quantify that this result is a consequence of providing a soft distillation of inductive biases (in this case translational equivariance). In Fig. 6, we explore the impact of applying a translation on the attention maps of a given layer. We observe that any translation of an object is reflected with a translation of the attention maps. Interestingly, this is unlike other methods, such as Deit-Ti⚗ , where the attention maps become messy after the translation of the original image. This observation suggests that their improvements may instead be attributed to some other factor, such as an implicit regularisation of the model.

Improved localisation of attention maps. We provide further insights into our orthogonal feature distillation method by analyzing the attention maps of various images. These attention maps show how well the model is attending to the salient objects in an image [7]. We compare with various other distillation methods and show the results in Fig. 7. We observe that the attention maps of our method are clustered around the boundaries of the objects, unlike the other two methods where the attention maps are spread over the entire image. In fact, we observe that our distilled model can attend much more to the salient object than the much larger CiT-S model.

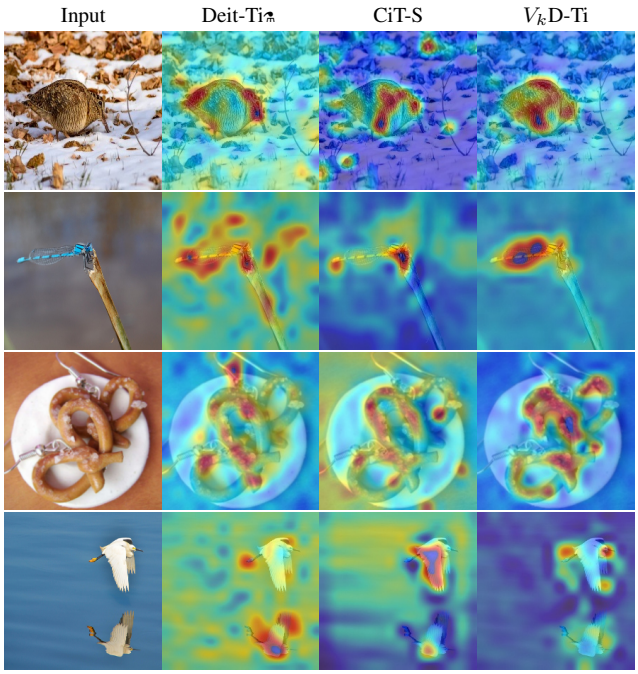


Figure 7. Qualitative comparison to other transformer distillation methods. The best channel is selected qualitatively for all examples shown. V_kD-Ti is compared against both the same size Deit-Ti and a much larger (3.7×) CiT-S. Best viewed in colour.

Architecture agnostic. Our method is agnostic in the choice of features and can be applied to various classifiers, object detectors, or generative models. For example, in section 4.1, we distill from a CNN to a transformer, in section 4.2, we distill between two transformers, and in section 4.3 we perform distillation in the other direction, from a transformer to a CNN.

5. Conclusion

In this work, we present a novel projection layer with a principled foundation centered on preserving the intra-batch feature similarity. The core idea of maintaining feature similarity ensures that the projection layer does not distort the underlying student representation, thus maximizing the knowledge transfer to the student backbone. We show that enforcing this constraint is equivalent to parameterising the projection weights to have orthonormal rows or columns. Our simple drop-in replacement for the projection layer leads to improved performance across a wide range of distillation tasks, from image classification to object detection. To further improve the generality of this framework, we show that whitening the teachers' features is sufficient and more effective in extending to generative tasks than other methods. We show in the experiments that our method improves state-of-the-art by up to 4.4% for image classification and 2.6% for object detection.

References

[1] Martin Arjovsky, Amar Shah, and Yoshua Bengio. Unitary evolution recurrent neural networks. *ICML*, 2016. 2

[2] Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. *ICLR*, 2022. 2

[3] Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. *CVPR*, 2022. 1, 2

[4] Kartikeya Bhardwaj, Naveen Suda, and Radu Marculescu. Dream distillation: A data-independent model compression framework. *ICML Joint Workshop on On-Device Machine Learning and Compact Deep Neural Network Representations (ODML-CDNNR)*, 2019. 1

[5] Daniel Brooks, Olivier Schwander, Frederic Barbaresco, Jean-Yves Schneider, and Matthieu Cord. Riemannian batch normalization for spd neural networks. *NeurIPS*, 2019. 2

[6] Adrian Bulat and Georgios Tzimiropoulos. XNOR-Net++: Improved Binary Neural Networks. *BMVC*, 2019. 1

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 8

[8] Liqun Chen, Dong Wang, Zhe Gan, Jingjing Liu, Ricardo Henao, and Lawrence Carin. Wasserstein Contrastive Representation Distillation. *CVPR*, 2020. 1, 2

[9] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling Knowledge via Knowledge Review. *CVPR*, 2021. 1, 2

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *ICML*, 2020. 2

[11] Xinlei Chen and Kaiming He. Exploring Simple Siamese Representation Learning. *CVPR*, 2021. 2

[12] Xianing Chen, Qiong Cao, Yujie Zhong, Jing Zhang, Shenghua Gao, and Dacheng Tao. Dearkd: Data-efficient early knowledge distillation for vision transformers. *CVPR*, 2022. 1, 5

[13] Yudong Chen, Sen Wang, Jiajun Liu, Xuwei Xu, Frank de Hoog, and Zi Huang. Improved Feature Distillation via Projector Ensemble. *NeurIPS*, 2022. 2, 7

[14] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. *ICCV*, 2019. 2

[15] Andrew Cotter, Joseph Keshet, and Nathan Srebro. Explicit approximations of the gaussian kernel. *arXiv preprint*, 2011. 3

[16] Kaiwen Cui, Yingchen Yu, Fangneng Zhan, Shengcai Liao, Shijian Lu1, and Eric Xing. Kd-dlgan: Data limited image generation via knowledge distillation. *CVPR*, 2023. 1, 4, 5, 6, 7

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ACL*, 2019. 1

[18] Alan Edelman, T. A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal of Matrix Analysis and Applications*, 1998. 4

[19] Mohamed Elfeki, Camille Couprie, Morgane Riviere, and Mohamed Elhoseiny. Gdpp: Learning diverse generations using determinantal point process. *ICML*, 2019. 4

[20] Gongfan Fang, Jie Song, Xinchao Wang, Chengchao Shen, Xingen Wang, and Mingli Song. Contrastive Model Inversion for Data-Free Knowledge Distillation. *IJCAI*, 2021. 1

[21] Spyros Gidaris, Andrei Bursuc, Gilles Puy, Nikos Komodakis, Matthieu Cord, and Patrick Pérez. Obow: Online bag-of-visual-words generation for self-supervised learning. *CVPR*, 2021. 2

[22] Jean Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent a new approach to self-supervised learning. *NeurIPS*, 2020. 2

[23] Jia Guo, Minghao Chen, Yao Hu, Chen Zhu, Xiaofei He, and Deng Cai. Reducing the Teacher-Student Gap via Spherical Knowledge Distillation. *arXiv preprint*, 2020. 2

[24] Shaopeng Guo, Yujie Wang, Quanquan Li, and Junjie Yan. DMCP: Differentiable Markov Channel Pruning for Neural Networks. *CVPR*, 2020. 1

[25] George Bruce Halsted. The collected mathematical papers of arthur cayley. *The American Mathematical Monthly*, 1899. 4

[26] Zhiwei Hao, Jianyuan Guo, Ding Jia, Kai Han, Yehui Tang, Chao Zhang, Han Hu, and Yunhe Wang. Learning efficient vision transformers via fine-grained manifold distillation. *NeurIPS*, 2022. 5

[27] Charles R. Harris, K. Jarrod Millman, Stefan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 2020. 4

[28] Bobby He and Mete Ozay. Feature Kernel Distillation. *ICLR*, 2022. 2

[29] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. *CVPR*, 2020. 2

[30] Kyle Helfrich, Devin Willmott, and Qiang Ye. Orthogonal recurrent neural networks with scaled cayley transform. *PMLR*, 2018. 2

[31] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. *ICCV*, 2019. 2

[32] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. *AAAI*, 2019. 2

[33] Nicholas J. Higham. The scaling and squaring method for the matrix exponential revisited. *SIAM Journal on Matrix Analysis and Applications*, 2005. 4

[34] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *NeurIPS*, 2015. 1, 2, 7

[35] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *ICLR*, 2019. 2

[36] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint*, 2021. 3

[37] Youmin Kim, Jinbae Park, Younho Jang, Muhammad Ali, and Tae-hyun Oh Sung-ho Bae. Distilling Global and Local Logits with Densely Connected Relations. *ICCV*, 2021. 2

[38] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *NeurIPS*, 2012. 1

[39] Yann Lecun. Optimal Brain Damage. *NeurIPS*, 1990. 1

[40] Xiaojie Li, Jianlong Wu, Hongyu Fang, Yue Liao, Fei Wang, and Chen Qian. Local correlation consistency for knowledge distillation. In *ECCV*, 2020. 2

[41] Mingbao Lin, Rongrong Ji, Yan Wang, Yichen Zhang, Baochang Zhang, Yonghong Tian, and Ling Shao. HRank: Filter Pruning using High-Rank Feature Map. *CVPR*, 2020. 1

[42] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context. *arXiv preprint*, 2015. 6

[43] Yufan Liu, Jiajiong Cao, Bing Li, Chunfeng Yuan, Weiming Hu, Yangxi Li, and Yunqiang Duan. Knowledge distillation via instance relationship graph. *CVPR*, 2019. 2

[44] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured Knowledge Distillation for Semantic Segmentation. *CVPR*, 2019. 1, 4

[45] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *ICLR*, 2016. 1

[46] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *ICLR*, 2019. 6

[47] Roy Miles and Krystian Mikolajczyk. Cascaded channel pruning using hierarchical self-distillation. *BMVC*, 2020. 1

[48] Roy Miles and Krystian Mikolajczyk. Understanding the role of the projector in knowledge distillation. *AAAI*, 2024. 5

[49] Roy Miles, Adrian Lopez Rodriguez, and Krystian Mikolajczyk. Information Theoretic Representation Distillation. *BMVC*, 2022. 1, 2

[50] Roy Miles, Mehmet Kerim Yucel, Bruno Manganelli, and Albert Saa-Garriga. MobileVOS: Real-Time Video Object Segmentation Contrastive Learning meets Knowledge Distillation. *CVPR*, 2023. 1, 2

[51] K L Navaneet, Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. SimReg: Regression as a Simple Yet Effective Tool for Self-supervised Knowledge Distillation. *BMVC*, 2021. 7

[52] Yulei Niu, Long Chen, Chang Zhou, and Hanwang Zhang. Respecting transfer gap in knowledge distillation. *NeurIPS*, 2022. 2

[53] Wonpyo Park, Kakao Corp, Dongju Kim, and Yan Lu. Relational Knowledge Distillation. *CVPR*, 2019. 1

[54] Nikolaos Passalis and Anastasios Tefas. Learning Deep Representations with Probabilistic Knowledge Transfer. *ECCV*, 2018. 7

[55] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. *NeurIPS 2017 Workshop Autodiff homepage*, 2017. 5

[56] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 2019. 4

[57] Baoyun Peng, Xiao Jin, Dongsheng Li, Shunfeng Zhou, Yichao Wu, Jiaheng Liu, Zhaoning Zhang, and Yu Liu. Correlation congruence for knowledge distillation. *CVPR*, 2019. 2

[58] Hendrik Purwins, Bo Li, Tuomas Virtanen, Jan Schluter, Shuo-Yiin Chang, and Tara Sainath. Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 2019. 1

[59] Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. *arXiv preprint*, 2023. 3

[60] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing Network Design Spaces. *CVPR*, 2020. 5

[61] Sucheng Ren, Zhengqi Gao, Tianyu Hua, Zihui Xue, Yonglong Tian, Shengfeng He, and Hang Zhao. Co-advise: Cross Inductive Bias Distillation. *CVPR*, 2022. 5

[62] Pierre H. Richemond, Jean-Bastien Grill, Florent Altché, Corentin Tallec, Florian Strub, Andrew Brock, Samuel Smith, Soham De, Razvan Pascanu, Bilal Piot, and Michal Valko. Byol works even without batch statistics. *arXiv preprint*, 2020. 2

[63] Pierre H. Richemond, Allison Tam, Yunhao Tang, Florian Strub, Bilal Piot, and Felix Hill. The edge of orthogonality: A simple view of what makes byol tick. *arXiv preprint*, 2023. 2

[64] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints For Thin Deep Nets. *ICLR*, 2015. 1, 2, 7

[65] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *NeurIPS Workshop on Energy Efficient Machine Learning and Cognitive Computing*, 2019. 2

[66] Seungwoo Son, Namhoon Lee, and Jaeho Lee. Maskedkd: Efficient distillation of vision transformers with masked images. *arXiv preprint*, 2023. 5

[67] Hwanjun Song, Deqing Sun, Sanghyuk Chun, Varun Jampani, Dongyoon Han, Byeongho Heo, Wonjae Kim, and

Ming-Hsuan Yang. Vidt: An efficient and effective fully transformer-based object detector. *ICLR*, 2021. 6

[68] Suraj Srinivas and François Fleuret. Knowledge transfer with jacobian matching. *ICML*, 2018. 2

[69] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *ICLR*, 2019. 1, 2

[70] Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised Learning Dynamics without Contrastive Pairs. *ICML*, 2021. 2

[71] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *PMLR*, 2021. 1, 5, 6

[72] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *ICCV*, 2021. 5

[73] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *ECCV*, 2022. 5

[74] Fred Tung and Greg Mori. Similarity-preserving knowledge distillation. *ICCV*, 2019. 2, 7

[75] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008. 4

[76] Xiang Wang, Xinlei Chen, Simon S. Du, and Yuandong Tian. Towards demystifying representation learning with non-contrastive self-supervision. *arXiv preprint*, 2022. 2

[77] Guodong Xu, Ziwei Liu, Xiaoxiao Li, and Chen Change Loy. Knowledge Distillation Meets Self-supervision. *ECCV*, 2020. 2

[78] Guodong Xu, Ziwei Liu, Xiaoxiao Li, and Chen Change Loy. Knowledge distillation meets self-supervision. *ECCV*, 2020. 2

[79] Jing Yang, Brais Martinez, Adrian Bulat, and Georgios Tzimiropoulos. Knowledge distillation via softmax regression representation learning. In *ICLR*, 2021. 2

[80] Jing Yang, Xiatian Zhu, Adrian Bulat, Brais Martinez, and Georgios Tzimiropoulos. Knowledge distillation meets open-set semi-supervised learning. *arXiv:2205.06701*, 2022. 2

[81] Ziqing Yang, Yiming Cui, Zhipeng Chen, Wanxiang Che, Ting Liu, Shijin Wang, and Guoping Hu. TextBrewer: An Open-Source Knowledge Distillation Toolkit for Natural Language Processing. *ACL*, 2020. 2

[82] Zhendong Yang, Ailing Zeng, Zhe Li, Tianke Zhang, Chun Yuan, and Yu Li. From knowledge distillation to self-knowledge distillation: A unified approach with normalized loss and customized soft labels. *ICCV*, 2023. 5

[83] Junho Yim. A Gift from Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning. *CVPR*, 2017. 2

[84] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. Mixup: Beyond empirical risk minimization. *ICLR*, 2017. 5

[85] Linfeng Zhang, Xin Chen, Xiaobing Tu, Pengfei Wan, Ning Xu, and Kaisheng Ma. Wavelet knowledge distillation: Towards efficient image-to-image translation. *CVPR*, 2022. 1

[86] Borui Zhao, Renjie Song, and Yiyu Qiu. Decoupled Knowledge Distillation. *CVPR*, 2022. 2

[87] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. *NeurIPS*, 2020. 7

[88] Shuchang Zhou, Yuxin Wu, Zekun Ni, Xinyu Zhou, He Wen, and Yuheng Zou. DoReFa-Net: Training Low Bitwidth Convolutional Neural Networks with Low Bitwidth Gradients. *arXiv preprint*, 2016. 1

[89] Jinguo Zhu, Shixiang Tang, Dapeng Chen, and Shijie Yu. Complementary Relation Contrastive Distillation. *CVPR*, 2021. 2