# Weight Copy and Low-Rank Adaptation for Few-Shot Distillation of Vision Transformers

Diana-Nicoleta Grigore[1,◇], Mariana-Iuliana Georgescu[1,◇], Jon Alvarez Justo[2], Tor Johansen[2], Andreea Iuliana Ionescu[3], and Radu Tudor Ionescu[1⋆]

[1] University of Bucharest, Romania
[2] Norwegian University of Science and Technology, Norway
[3] University of Medicine and Pharmacy Carol Davila, Romania

**Abstract.** Few-shot knowledge distillation recently emerged as a viable approach to harness the knowledge of large-scale pre-trained models, using limited data and computational resources. In this paper, we propose a novel few-shot feature distillation approach for vision transformers. Our approach is based on two key steps. Leveraging the fact that vision transformers have a consistent depth-wise structure, we first copy the weights from intermittent layers of existing pre-trained vision transformers (teachers) into shallower architectures (students), where the intermittence factor controls the complexity of the student transformer with respect to its teacher. Next, we employ an enhanced version of Low-Rank Adaptation (LoRA) to distill knowledge into the student in a few-shot scenario, aiming to recover the information processing carried out by the skipped teacher layers. We present comprehensive experiments with supervised and self-supervised transformers as teachers, on five data sets from various domains, including natural, medical and satellite images. The empirical results confirm the superiority of our approach over competitive baselines. Moreover, the ablation results demonstrate the usefulness of each component of the proposed pipeline.

**Keywords:** Knowledge Distillation · Low Rank Adaptation · Vision Transformers · Few-Shot Distillation

## 1 Introduction

Vision transformers [4, 19, 32, 45, 54, 69] have revolutionized the computer vision research in the past few years, reaching state-of-the-art performance across a broad range of tasks, such as object recognition [19, 45, 49, 69], object detection [7, 40, 71, 78, 80], image segmentation [10, 11, 20, 23, 73], image translation [33, 56, 61], among many others [32]. Since the accuracy tends to grow as the model gets larger [19, 49], most of the attention has been dedicated to building larger and more powerful models. However, the typically large size and slow inference speed of transformer-based architectures hinders the deployment of such models

⋆ Corresponding author: raducu.ionescu@gmail.com. ◇ Equal contribution.

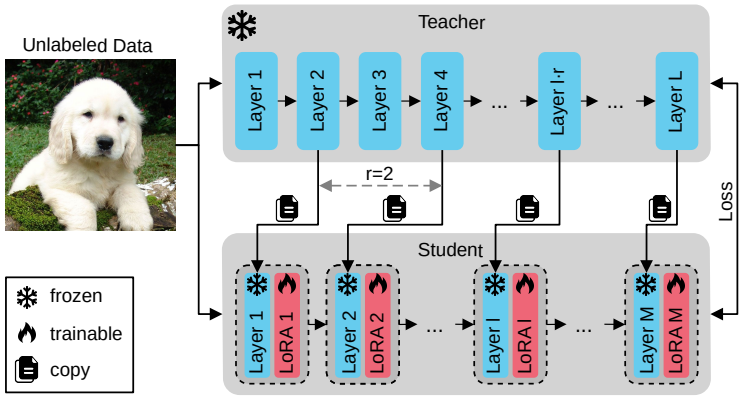


**Fig. 1:** Our feature distillation framework is based on two steps. In the first step, weights from intermittent layers of the teacher transformer are copied directly into the student, where the intermittence factor $r$ coincides with the compression ratio between the teacher and the student transformers. In the second step, enhanced LoRA blocks are integrated into the student network. The enhanced LoRA blocks are trained via feature distillation on unlabeled images. In the illustrated example, the compression ratio is $r = 2$.

on environments with limited computational resources, *e.g.* on edge or mobile devices.

The success of vision transformers lies in the "pre-training then fine-tuning" paradigm, which originates from the natural language processing domain [16,66]. The multi-head attention layers inside transformers can capture long-range spatial relationships among tokens, but this flexibility has a significant downside: transformers can easily overfit small training sets and experience poor generalization capabilities [37,44]. Unlike convolutional nets, which benefit from the inductive bias of convolutional filters with small receptive fields [35], transformers requires huge amounts of data in the pre-training stage to avoid overfitting [19], regardless of the chosen supervised [19,45,69] or self-supervised [18,21,25,38,60,67] training setup. Therefore, to efficiently train large-scale transformer models, powerful and expensive machines are necessary, which are not widely available to researchers. Moreover, large amounts of data are not always available in some domains, *e.g.* hyperspectral image segmentation [30].

To mitigate the challenges of training large-scale models on large amounts of data on machines with limited computational resources, researchers have proposed the few-shot knowledge distillation (FSKD) paradigm [24,39,51,53,55,58, 59, 76, 79], which was explored in both language [53, 58, 79] and vision [24, 39, 41, 42, 51, 59, 76] domains. This paradigm allows the trained model (called student) to benefit from the knowledge learned by large-scale transformers (called teachers), while significantly reducing the training time and data set size.

We emphasize that FSKD has not been extensively explored in the vision domain [24, 39, 41, 42, 51, 55, 59, 76], with even less studies focused on the pre-training stage of vision transformers [42]. To this end, we propose a novel few-

shot feature distillation approach for vision transformers based on intermittent **We**ight **Co**pying and **Lo**w-**R**ank **A**daptation (WeCoLoRA), as illustrated in Figure 1. Our approach is divided into two steps. For the first step, we leverage the fact that vision transformers have a consistent depth-wise structure, *i.e.* the input and output dimensions are compatible across transformer blocks. This allows us to directly copy the weights from intermittent layers of existing pre-trained vision transformers (teachers) into shallower architectures (students). Here, we use the intermittence factor to control the complexity (size) of the student transformer with respect to its teacher. In the second step, we employ an enhanced version of Low-Rank Adaptation (LoRA) [28] to distill knowledge into the student in a few-shot scenario, aiming to recover the information processing carried out by the teacher layers that were skipped in the first step.

We perform the pre-training stage of efficient student transformers via few-shot knowledge distillation on various subsets of ImageNet-1K [15]. Then, we carry out linear probing experiments on five downstream data sets from different image domains, namely ImageNet-1K [15], CIFAR-100 [34], ChestX-ray14 [68], iNaturalist [64] and RESISC45 [12]. We compare WeCoLoRA with state-of-the-art competitors, namely DMAE [3] and DeiT [62], showing that our approach leads to superior results. Furthermore, we conduct an ablation study to demonstrate that each proposed component brings significant performance gains. Additionally, we perform an analysis of the distilled features, which explains why WeCoLoRA produces more robust features than competing methods.

In summary, our contribution is threefold:

- We propose a novel few-shot feature distillation approach for vision transformers based on (*i*) intermittent weight copying and (*ii*) enhanced low-rank adaption, called WeCoLoRA.
- We present few-shot and linear probing experiments on five benchmark data sets comprising natural, medical and satellite images, demonstrating the utility of our training pipeline across different domains.
- We analyze the features learned by our distilled models in comparison with those of the strongest competitor, showing that our approach generates more robust and discriminative features.

## 2 Related Work

**Knowledge distillation.** Knowledge distillation (KD) [27, 47, 50, 57], a.k.a. teacher-student training, is an efficiency-boosting technique meant to reduce the computational load that comes with large models. It emerged [47] from the union of model compression [2, 27] and learning under privileged information [65]. Our study is only preoccupied with the former task, being aimed at proposing an approach that involves transferring expertise from a heavy teacher model to a lighter student model [6, 63], with the latter finally being able to learn more discriminative features than through a conventional training procedure [13, 14, 29].

**KD for vision transformers.** KD was originally applied to vision transformers by Touvron *et al.* [62], who proposed Data-Efficient Image Transformers (DeiT)

based on leveraging attention as knowledge and enhancing the student's capabilities through a learned distillation token. Since then, several other studies used KD on vision transformers [22, 31, 42, 70, 77]. Close to our work, MiniViT [74] uses a parameter reduction technique that involves copying weights (and applying slight perturbations) between the teacher and a smaller student transformer, followed by distillation through self-attention. In comparison, after copying the weights, we employ an enhanced version of LoRA [28] during the information recovery process, which leads to significantly better results in the few-shot scenario. Recently, distilling Masked Autoencoders [25] has gathered some traction [3, 36, 42], with most approaches achieving data efficiency through masking very large percentages (up to 98%) of the input. Our data reduction strategy comes solely from a reduction of the number of unlabeled samples, being more fit for cases where data scarcity is the issue.

**Few-shot knowledge distillation.** Zero-shot knowledge distillation [5, 8, 46, 75, 82] implies that the teacher and student do not share the same data sources. As a consequence, the latter should obtain its knowledge through synthetic examples [9, 43]. Black-box KD [5, 51] involves generating synthetic images from the teacher's training distribution, along with their corresponding labels, offering the dual benefit of privacy preservation and effective learning. Zero-shot KD emerged as a solution to the lack of knowledge about the training data used by the teacher, but it suffers from the problem of collapse [17], caused by similar training samples. A less strict alternative that mitigates this issue is to employ few-shot KD [24, 39, 41, 42, 51, 55, 59, 76]. To perform FSKD, Li *et al.* [39] introduced an additional $1 \times 1$ convolutional layer at the end of each block of a CNN to support recovering the abilities of a teacher, using a small set of unlabeled examples. To the best of our knowledge, we are the first to employ few-shot KD in the pre-training stage of lightweight vision transformers.

**Low-Rank Adaptation.** LoRA and its variations [28, 72, 81] represent parallel performance improvement methods, enabling the use of powerful large models through lightweight fine-tuning. Consequently, LoRA reduces the inference cost and the requirement to use large data sets. Following other lines of work, the authors of LoRA [28] showed that over-parameterized models reside on a very low internal dimension when making their decisions [52]. The proven hypothesis is that the change in weights during domain adaptation or fine-tuning should also be happening at a low intrinsic rank. To boost training efficiency, we thus combine Low-Rank Adaptation [28] and KD [27] in our study. To the best of our knowledge, there is no prior work that employs weight copying and LoRA as a method for few-shot feature distillation in an unsupervised setting.

## 3 Method

### 3.1 WeCoLoRA Architecture

Our knowledge distillation framework, WeCoLoRA, is formally described in Algorithm 1. Our method follows the knowledge distillation paradigm of compressing a larger model into a smaller one. Therefore, our goal is to obtain a student

**S** which runs faster than the teacher **T** during inference. However, we also use significantly less training data during the knowledge distillation process, since it is not always feasible to obtain the same amount of data as the teacher was (pre-)trained on. Our knowledge distillation method is designed for vision transformers [19]. Given a pre-trained transformer-based teacher **T** with $L$ layers, we create the student **S** that has $\lfloor \frac{L}{r} \rfloor$ layers, where $r$ is the reduction ratio of the number of layers and $\lfloor \cdot \rfloor$ is the floor approximation function. The first step of our algorithm, which is described in steps 2-4 of Algorithm 1, is to copy the pre-trained weights of **T** to **S**. Since the student **S** has fewer layers than the teacher, due to the reduction ratio $r$, the weights from the layer at index $r \cdot l$ of the teacher are copied to the layer at index $l$ of the student, where $l \in \{1, 2, ..., \lfloor \frac{L}{r} \rfloor\}$.

After the intermittent weight copying step, we apply an adaptor layer, such that the student model is able to recover the weights which are not transferred from the teacher. In this way, we aim to minimize the performance gap between the teacher and the student. We employ the LoRA framework [48] as the adaptor, due to its competitive performance obtained in model adaptation. LoRA can be applied to any fully-connected layer, which makes it a great choice for the transformer architecture. Therefore, for a pre-trained fully-connected layer $f$, with the weight matrix $\boldsymbol{W} \in \mathbb{R}^{d_1 \times d_2}$, LoRA learns a matrix $\boldsymbol{B} \in \mathbb{R}^{d_1 \times k}$ and a matrix $\boldsymbol{A} \in \mathbb{R}^{k \times d_2}$, where $k$ is the matrix rank, $k \leq min(d_1, d_2)$, and $d_1$ and $d_2$ are the matrix dimensions. During the knowledge distillation process, $\boldsymbol{W}$ is kept frozen, so only the parameters $\boldsymbol{A}$ and $\boldsymbol{B}$ are updated. The output of the layer $f$ becomes $f(x) = \boldsymbol{W} \cdot x + \boldsymbol{B} \cdot \boldsymbol{A} \cdot x$, where $x \in \mathbb{R}^{d_1}$ is the input and $f(x) \in \mathbb{R}^{d_1 \times d_2}$.

Luo *et al.* [48] proposed to apply LoRA only to the projection matrices corresponding to the query and value tokens $\boldsymbol{Q}$ and $\boldsymbol{V}$, respectively. However, in order to fully replicate the functionality of the teacher, we conjecture that it is better to apply LoRA to each component of the transformer block. Therefore, for each head $h_i$, $\forall i \in \{1, 2, ..., H\}$ of the multi-head self-attention (MSA) layer, the outputs of the query $\boldsymbol{Q}_i$, value $\boldsymbol{V}_i$ and key $\boldsymbol{K}_i$ applied to the input $x$ become:

$$\begin{aligned} \boldsymbol{Q}_i &= \boldsymbol{W}_{Q_i} \cdot x + \boldsymbol{B}_{Q_i} \cdot \boldsymbol{A}_{Q_i} \cdot x, \\ \boldsymbol{K}_i &= \boldsymbol{W}_{K_i} \cdot x + \boldsymbol{B}_{K_i} \cdot \boldsymbol{A}_{K_i} \cdot x, \\ \boldsymbol{V}_i &= \boldsymbol{W}_{V_i} \cdot x + \boldsymbol{B}_{V_i} \cdot \boldsymbol{A}_{V_i} \cdot x, \end{aligned} \tag{1}$$

where $\boldsymbol{W}_{Q_i}$, $\boldsymbol{W}_{K_i}$ and $\boldsymbol{W}_{V_i}$ are the query, key and value projection matrices, while $\boldsymbol{B}_{*_i}$ and $\boldsymbol{A}_{*_i}$ are the matrices learned by LoRA, where $* \in \{Q, K, V\}$.

The formula to compute the output of the head $h_i$ remains unaltered, being equal to $h_i = \mathit{Attention}(\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i)$, $\forall i \in \{1, 2, ..., H\}$, where:

$$\mathit{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q} \cdot \boldsymbol{K}^\top}{\sqrt{d_k}}\right) \cdot \boldsymbol{V}, \tag{2}$$

and $d_k$ is the dimension of $\boldsymbol{K}$.

**Algorithm 1:** Knowledge Distillation with WeCoLoRA

**Input:** $\mathcal{D} = \{\boldsymbol{I}_1, \boldsymbol{I}_2, .., \boldsymbol{I}_n\}$ - the training set of unlabeled images, $\mathbf{T}$ - the pre-trained teacher model, $\theta_{\mathbf{T}}$ - pre-trained weights of the teacher, $\eta$ - the learning rate, $r$ - the depth reduction factor, $k$ - the rank of the low-rank matrices, $L$ - the number of layers of the teacher $\mathbf{T}$.

**Output:** $\theta_{\mathbf{S}}$ - the trained weights of the student model.

$n \leftarrow |\mathcal{D}|$; ◁ get the number of training samples
$M \leftarrow \lfloor \frac{L}{r} \rfloor$; ◁ compute the number of layers of the student $\mathbf{S}$
**foreach** $l \in \{1, 2, ..., M\}$ **do**
  $\theta_{\mathbf{S}_l} \leftarrow \theta_{\mathbf{T}_{l \cdot r}}$; ◁ copy weights from the teacher
$\theta^+_{\mathbf{S}} \leftarrow \emptyset$; ◁ initialize trainable weights
**foreach** $l \in \{1, 2, ..., M\}$ **do**
  $\mathbf{S}_l \leftarrow \mathbf{S}_l \cup \text{WeCoLoRA}(\mathbf{S}_l)$; ◁ add adaptor, as described in Section 3.1
  $\theta^+_{\mathbf{S}_l} \leftarrow \theta^+_{\mathbf{S}_l} \cup \{\boldsymbol{A}_{*_l}, \boldsymbol{B}_{*_l}\}$; ◁ add the new weights
**repeat**
  **foreach** $i \in \{1, 2, ..., n\}$ **do**
    $\boldsymbol{E}^{\mathbf{T}}_i \leftarrow \mathbf{T}(\boldsymbol{I}_i, \theta_{\mathbf{T}})$; ◁ get embedding from teacher
    $\boldsymbol{E}^{\mathbf{S}}_i \leftarrow \mathbf{S}(\boldsymbol{I}_i, \theta_{\mathbf{S}} \cup \theta^+_{\mathbf{S}})$; ◁ get embedding from student
    $\mathcal{L}_{\text{KD}} \leftarrow \mathcal{L}(\boldsymbol{E}^{\mathbf{T}}_i, \boldsymbol{E}^{\mathbf{S}}_i)$; ◁ apply Eq. (5)
    $\theta^+_{\mathbf{S}} \leftarrow \theta^+_{\mathbf{S}} - \eta \cdot \nabla \mathcal{L}_{\text{KD}}$; ◁ update only the newly added weights
**until** *convergence*;
$\theta_{\mathbf{S}} \leftarrow \theta_{\mathbf{S}} \oplus \theta^+_{\mathbf{S}}$; ◁ integrate trained weights into the copied weights

We also add adaptation parameters to the output projection layer in the multi-head attention, resulting in:

$$MultiHead(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \boldsymbol{W}_O \cdot [h_1, h_2, .., h_H] + \boldsymbol{B}_O \cdot \boldsymbol{A}_O \cdot [h_1, h_2, .., h_H], \quad (3)$$

where $[\cdot]$ is the concatenation operation, $\boldsymbol{W}_O$ is the output projection matrix, $\boldsymbol{B}_O$ and $\boldsymbol{A}_O$ are the parameters introduced by LoRA, and $H$ is the number of heads.

Further, we also integrate LoRA into the feed-forward networks (FFNs). Therefore, the output of an FFN from a transformer block is:

$$\begin{aligned} \text{FFN}(x) &= \boldsymbol{W}_{f_2} \cdot \text{FFL}(x) + \boldsymbol{B}_{f_2} \cdot \boldsymbol{A}_{f_2} \cdot \text{FFL}(x), \\ \text{FFL}(x) &= \sigma(\boldsymbol{W}_{f_1} \cdot x + \boldsymbol{B}_{f_1} \cdot \boldsymbol{A}_{f_1} \cdot x), \end{aligned} \quad (4)$$

where FFL is the output of the first layer in the FFN model, $\boldsymbol{W}_{f_1}$ and $\boldsymbol{W}_{f_2}$ are the parameters of the feed-forward network, $\sigma$ is the activation function, and $\boldsymbol{A}_*$ and $\boldsymbol{B}_*$ are the matrices learned by LoRA, where $* \in \{f_1, f_2\}$. The biases in the FFN network were omitted to enhance the clarity of the presentation. By introducing LoRA in every component of the transformer block, we obtain an enhanced version of LoRA-based transformer blocks.

### 3.2 Knowledge Distillation with WeCoLoRA

Our distillation method does not require training labels, therefore, the knowledge distillation data set is $\mathcal{D} = \{\boldsymbol{I}_1, \boldsymbol{I}_2, .., \boldsymbol{I}_n\}$, where $\boldsymbol{I}_i$ is an image sample and $n$ is the number of samples in the data set. This is because the knowledge distillation procedure is performed in the latent space. Moreover, we chose $n$ to be typically small, resulting in a few-shot feature distillation training setup.

The goal of applying knowledge distillation is to transfer the knowledge of the pre-trained teacher $\mathbf{T}$ into the LoRA-enhanced student $\mathbf{S}$. Following Luo *et al.* [48], during optimization, we only update the newly added weights of the student $\mathbf{S}$, denoted as $\theta_{\mathbf{S}}^{+} = \{\boldsymbol{A}_*, \boldsymbol{B}_*\}$, where $* \in \{Q_l, K_l, V_l, O_l, f_{l,1}, f_{l,2}\}$, $\forall l \in \{1, 2, ..., \lfloor \frac{L}{r} \rfloor\}$. The operation of inserting trainable weights into the student is formally described in steps 6-8 of Algorithm 1.

In order to optimize the parameters $\theta_{\mathbf{S}}^{+}$ of the student, we minimize the absolute difference between the embedding $\boldsymbol{E}_i^{\mathbf{T}} \in \mathbb{R}^{t \times d}$ returned by the teacher $\mathbf{T}$ for image $\boldsymbol{I}_i$, and the embedding $\boldsymbol{E}_i^{\mathbf{S}} \in \mathbb{R}^{t \times d}$ returned by the student $\mathbf{S}$ for the same image $\boldsymbol{I}_i$, where $t$ is the number of tokens and $d$ is their hidden dimension. Formally, the proposed feature distillation is expressed as follows:

$$\mathcal{L}(\boldsymbol{E}_i^{\mathbf{T}}, \boldsymbol{E}_i^{\mathbf{S}}) = \left| \boldsymbol{E}_i^{\mathbf{T}} - \boldsymbol{E}_i^{\mathbf{S}} \right|, \forall i \in \{1, 2, ..., n\}, \tag{5}$$

where $n$ is the number of training images. This procedure is described in steps 10-14 of Algorithm 1.

After optimizing the parameters $\theta_{\mathbf{S}}^{+}$, we integrate them into the pre-trained parameters $\theta_{\mathbf{S}}$ (copied from the teacher $\mathbf{T}$) in step 16, as described by Luo *et al.* [48], in order to avoid affecting the running time during inference.

## 4 Experiments

We use ImageNet-1K [15] to perform knowledge distillation in various few-shot setups, considering subsets ranging between 1% and 10% of the number of samples in the original training set. We test the models via linear probing, using the official ImageNet-1K evaluation set. We also perform linear probing experiments on five downstream data sets: ImageNet-1K [15], CIFAR-100 [34], ChestX-ray14 [68], iNaturalist [64] and RESISC45 [12].

### 4.1 Data Sets

**ImageNet-1K**. In our self-supervised knowledge distillation process, we operate on ImageNet-1K [15], a large-scale data set with 1,281,167 color images from various sources, each having an average shape of $469 \times 387$ pixels. The collection covers 1,000 object categories, such as plants, vehicles, and everyday items.
**CIFAR-100**. CIFAR-100 [34] is a balanced data set of 60,000 images representing various visual entities, grouped in 100 classes. The categories are arranged in a hierarchical structure, under 20 superclasses. The official test set contains 10,000 pictures. Each color image has $32 \times 32$ pixels.

**ChestX-ray14**. The CXR14 [68] data set is composed of 112,120 frontal-view X-Ray images from 30,805 patients and depicts a collection of 14 common pathologies, as well as normal lung scans. Its annotations are binary multi-labels, indicating whether each of 14 lung diseases is present or not. The labels are based on radiological reports. The resolution of each image is $1024 \times 1024$. To evaluate our models, we use the test set comprising 20% of the images, on which we perform multi-label classification.

**iNaturalist**. iNaturalist [64] is a heavily imbalanced data set, proposing the fine-grained classification of 1,010 plant and animal species. Comprising a total size of 268,243 images with varying resolutions, its 2019 version offers an evaluation set of 3030 images.

**RESISC45**. NWPU-RESISC45 [12] incorporates 31,500 3-channel images from the aerial domain, which can be used for remote sensing classification. The data set comprises 45 scene classes and various spatial resolutions, where one pixel can represent a surface ranging from 20cm to 30m. In our evaluation setting, we use 80% of the data as the test set.

### 4.2 Implementation Details

We implement our method in Pytorch. We create the enhanced version of LoRA starting from the official code of LoRA[4] and extending it, in order to be applied to each component of the transformer layer. To demonstrate the generalization of our method to different teachers, we perform experiments with a supervised and a self-supervised teacher, respectively. As the supervised teacher backbone, we employ the ViT-B model [19] trained on ImageNet, available in the `timm`[5] library. The selected self-supervised model is also based on ViT-B, but pre-trained using the Masked Autoencoder framework [25]. Since the goal is to create a model which is faster and lighter, we refrain from using larger teacher models, which would be inefficient for our purpose.

To show that our method generalizes to different depth reduction factors, we perform experiments with $r = 2$ and $r = 3$. Since the number of layers $L$ in the teacher architecture (ViT-B) is 12, the number of layers in the student architecture becomes $M = 6$ (when $r = 2$) or $M = 4$ (when $r = 3$), respectively. We select the rank of the low-rank matrices from $k \in \{128, 256, 512\}$, which controls the number of parameters used to recover the skipped layers from the teacher.

We perform the knowledge distillation procedure for only 10 epochs. We set the learning rate to $10^{-3}$ when the compression factor $r$ equals 2, and to $10^{-4}$ when $r$ equals 3. The batch size is set to 128, with a gradient accumulation factor of 8. We perform weak data augmentation, similar to He *et al.* [25].

To demonstrate that our method transfers strong features, we only perform linear probing on the downstream tasks. For each downstream data set, we train the linear classifier using the best available linear probing training recipe. We also provide the code for an easier reproduction of the experiments. For a fair

[4] https://github.com/microsoft/LoRA
[5] https://pypi.org/project/timm

**Table 1:** Ablation results in terms of accuracy (in percentages) on ImageNet-1K [15], iNaturalist19 [64] and CIFAR-100 [34]. The results are obtained using a compression factor of $r = 2$, employing the self-supervised teacher ViT-B [25]. The parameter $\alpha$ represents the percentage of the original ImageNet-1K training set [15] used during knowledge distillation. The best accuracy on each downstream data set and each value of $\alpha$ is highlighted in bold.

| $\alpha$ | Weight Copy | LoRA | Enhanced LoRA | ImageNet | iNaturalist | CIFAR-100 |
|---|---|---|---|---|---|---|
| $\alpha = 1\%$ | ✗ | ✗ | ✗ | 3.5 | 2.4 | 9.1 |
| | ✓ | ✗ | ✗ | 30.2 | 19.2 | 32.1 |
| | ✓ | ✓ | ✗ | 32.4 | 21.6 | 30.7 |
| | ✓ | ✗ | ✓ | **36.8** | **22.7** | **33.3** |
| $\alpha = 10\%$ | ✗ | ✗ | ✗ | 15.8 | 14.5 | 28.9 |
| | ✓ | ✗ | ✗ | 53.7 | 31.2 | 48.2 |
| | ✓ | ✓ | ✗ | 33.0 | 28.4 | 41.1 |
| | ✓ | ✗ | ✓ | **56.0** | **34.5** | **49.8** |

comparison, we use the same linear probing training setting for all the models included in the comparison.

### 4.3 Results

**Ablation study.** We compare our method with various ablated versions and report the results in Table 1. We start by comparing our method to the conventional KD method based on training a student vision transformer from scratch, using knowledge distillation (first and fifth rows). This approach, which is well-known and widely-employed, does not perform well in the few-shot setting, since there is not enough data to learn the underlying distribution. Next, we perform experiments using an ablation version of our method, which copies the weights from the teacher (second and sixth rows). This approach, which we refer to as Weight Copying + Knowledge Distillation (WeCo+KD), updates all the weights of the student. Even though this is a very strong baseline, reaching an accuracy of 53.7% in the linear probing setting on ImageNet when 10% of the initial training data is used during knowledge distillation, it is still 2.3% behind our approach. Nevertheless, weight copying brings significant performance boosts over the standard distillation, confirming its practical utility.

The third ablated model uses weight copying and adds LoRA to the projection matrices corresponding to the query and value tokens $\boldsymbol{Q}$ and $\boldsymbol{V}$, following Luo *et al.* [48]. This approach performs poorly, reaching only 33.0% in terms of accuracy on the ImageNet downstream task, when 10% of the initial training data is employed. We conjecture that LoRA adds an insufficient number of learnable weights to properly learn the skipped layers of the teacher. Our full framework, WeCoLoRA, replaces standard LoRA with our enhanced version of LoRA, which is based on adding LoRA layers to each component of the transformer blocks. Our enhanced LoRA (fourth and eight rows in Table 1) outperforms both LoRA and WeCo+KD on all datasets. In summary, the ablation study demonstrates the utility of our novel components.

**Table 2:** Accuracy rates (in percentages) obtained on ImageNet-1K [15], iNaturalist19 [64] and CIFAR-100 [34] by two state-of-the-art knowledge distillation methods, DMAE [3] and DeiT [62], in comparison with WeCo+KD and WeCoLoRA. The results are obtained using a compression factor of $r = 2$, employing the self-supervised teacher ViT-B [25]. The parameter $\alpha$ represents the percentage of the original ImageNet-1K training set [15] used during knowledge distillation. The best accuracy on each downstream data set and each value of $\alpha$ is highlighted in bold.

| $\alpha$ | Method | Venue | ImageNet | iNaturalist | CIFAR-100 |
|---|---|---|---|---|---|
| $\alpha = 1$ | DMAE [3] | CVPR 2023 | 15.5 | 8.5 | 22.4 |
| | DeiT [62] | ICML 2021 | 16.5 | 12.9 | 28.8 |
| | WeCo+KD (ours, ablated) | - | 30.2 | 19.2 | 32.1 |
| | WeCoLoRA (ours) | - | **36.8** | **22.7** | **33.3** |
| $\alpha = 10$ | DMAE [3] | CVPR 2023 | 30.0 | 17.1 | 31.7 |
| | DeiT [62] | ICML 2021 | 39.9 | 29.2 | 48.1 |
| | WeCo+KD (ours, ablated) | - | 53.7 | 31.2 | 48.2 |
| | WeCoLoRA (ours) | - | **56.0** | **34.5** | **49.8** |
| - | ViT-B [25] (teacher) | CVPR 2022 | 66.1 | 41.6 | 55.1 |

**Comparison with state-of-the-art KD techniques.** We compare WeCoLoRA and WeCo+KD with two state-of-the-art knowledge distillation techniques, namely DMAE [3] and DeiT [62], in Table 5. We also report the performance obtained by the self-supervised teacher ViT-B [25] on each downstream task, as an upper bound for the few-shot methods. Interestingly, neither state-of-the-art method [3, 62] outperforms the ablated version of WeCoLoRA, namely WeCo+KD. This indicates that current methods, such as DMAE and DeiT, are not able to handle the limited amount of data that is typical to the few-shot setting. WeCoLoRA surpasses the two state-of-the-art KD methods [3, 62] by significant margins, *e.g.* the differences are between 16% and 26% on ImageNet-1K. Moreover, on the CIFAR-100 data set, our method obtains an accuracy that is only 5.3% below the teacher, while using half the number of layers and only using 10% of the training data during distillation. Since WeCo+KD, the ablated version of our method, surpasses both DMAE [3] and DeiT [62], we choose it as a strong baseline for the subsequent experiments.

**Results with multiple teachers and compression rates.** We report additional results with WeCoLoRA and its strong ablated version, WeCo+KD, in Tables 3 and 4. We alternatively employ two teachers in these experiments, a supervised ViT-B [19] and a self-supervised ViT-B [25]. We also consider two compression factors, namely $r = 2$ and $r = 3$, on five downstream tasks.

In Table 3, we report results when only 1% of the ImageNet-1K training set [15] is employed during the knowledge distillation procedure. In this scenario, WeCoLoRA outperforms the baseline in most cases. Remarkably, for $r = 2$, WeCoLoRA achieves an accuracy of 60.5% on ImageNet using only 1% of the training data, surpassing the accuracy of WeCo+KD by 25%.

We present results for 10% of the distillation data in Table 4. Once again, WeCoLoRA outperforms the baseline in most settings. However, the gains in favor of our method when using 10% of the data (Table 4) are generally lower

**Table 3:** Results of WeCoLoRA and WeCo+KD in terms of accuracy (in percentages) on ImageNet-1K [15], iNaturalist [64], NWPU-RESISC45 [12] and CIFAR-100 [34], and in terms of mean Area Under the Curve (AUC, in percentages) on ChestX-ray14 [68]. Results are reported for two teachers: supervised ViT-B [19] and self-supervised (SSL) ViT-B [25]. During the distillation procedure, only **1**% of the ImageNet-1K training set [15] is used. The best score on each data set for each teacher and each compression rate is highlighted in bold.

| Teacher | Compression factor $r$ | Distillation method | ImageNet | ChestX-ray14 | iNaturalist | RESISC45 | CIFAR-100 |
|---|---|---|---|---|---|---|---|
| ViT-B [19] (supervised) | 2 | WeCo+KD | 35.5 | 67.6 | 25.4 | 58.0 | 34.8 |
| | | WeCoLoRA | **60.5** | **69.5** | **43.4** | **67.5** | **60.6** |
| | 3 | WeCo+KD | 7.6 | 57.2 | 6.0 | 33.3 | 11.0 |
| | | WeCoLoRA | **35.8** | **67.6** | **26.6** | **58.8** | **36.5** |
| ViT-B [25] (SSL) | 2 | WeCo+KD | 30.2 | 65.7 | 19.2 | 53.8 | 32.1 |
| | | WeCoLoRA | **36.8** | **66.3** | **22.7** | **56.8** | **33.4** |
| | 3 | WeCo+KD | 32.4 | **66.7** | **21.1** | 42.2 | 34.0 |
| | | WeCoLoRA | **32.9** | 66.6 | **21.1** | **55.6** | **34.4** |

**Table 4:** Results of WeCoLoRA and WeCo+KD in terms of accuracy (in percentages) on ImageNet-1K [15], iNaturalist [64], NWPU-RESISC45 [12] and CIFAR-100 [34], and in terms of mean Area Under the Curve (AUC, in percentages) on ChestX-ray14 [68]. Results are reported for two teachers: supervised ViT-B [19] and self-supervised (SSL) ViT-B [25]. During the distillation procedure, only **10**% of the ImageNet-1K training set [15] is used. The best score on each data set for each teacher and each compression rate is highlighted in bold.

| Teacher | Compression factor $r$ | Distillation method | ImageNet | ChestX-ray14 | iNaturalist | RESISC45 | CIFAR-100 |
|---|---|---|---|---|---|---|---|
| ViT-B [19] (supervised) | 2 | WeCo+KD | 68.3 | 69.8 | 45.7 | **74.0** | 64.4 |
| | | WeCoLoRA | **69.2** | **70.0** | **49.5** | 70.5 | **68.3** |
| | 3 | WeCo+KD | 58.1 | 68.4 | 38.3 | **67.7** | 51.4 |
| | | WeCoLoRA | **58.3** | **69.1** | **41.6** | 66.8 | **54.1** |
| ViT-B [25] (SSL) | 2 | WeCo+KD | 53.7 | 69.7 | 31.2 | 61.4 | 48.2 |
| | | WeCoLoRA | **56.0** | **69.8** | **34.5** | **62.3** | **49.8** |
| | 3 | WeCo+KD | 40.6 | **67.6** | 22.7 | 53.5 | 34.6 |
| | | WeCoLoRA | **41.0** | 67.3 | **23.5** | **60.2** | **39.5** |

than distilling on 1% of the data (Table 3). We consider that 10% of ImageNet-1K, *i.e.* about 100K training images, is at the upper end of the scenarios that can still be regarded as *few-shot*. The higher number of trainable parameters in WeCo+KD, combined with the large training set, allow it to recover the gap with respect to WeCoLoRA. To further demonstrate the superiority of WeCoLoRA

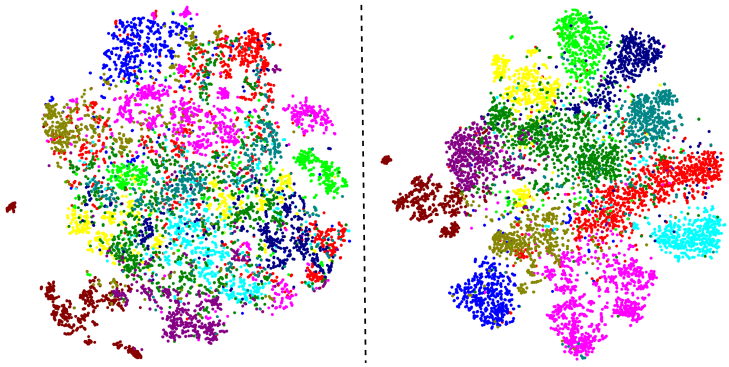

**Fig. 2:** Visualizations of the latent spaces learned by WeCo+KD (on the left-hand side) and our method (on the right-hand side). Both visualizations are obtained with t-SNE. The embeddings correspond to images from the RESISC45 [12] test set, before linear probing. The features are extracted from student models that are distilled from the supervised teacher ViT-B [25], using a compression factor of $r = 2$. During the distillation procedure, only 1% of ImageNet data set [15] is used. The colors correspond to the class labels from RESISC45. Best viewed in color.

in other few-shot scenarios, we present additional results when using 2% and 5% of the data during distillation, in the supplementary.

**Latent space visualizations.** In Figure 2, we illustrate t-SNE visualizations of the embeddings learned with the proposed method (WeCoLoRA) versus its ablated version denoted as WeCo+KD. The embeddings are obtained on the RESISC45 [12] test set. We illustrate the embeddings of the student obtained through distilling the supervised ViT-B [19] teacher, with a compression ratio set to $r = 2$. We notice that our WeCoLoRA is able to learn a discriminative feature distribution during distillation, disentangling the out-of-domain (downstream) data samples from RESISC45 into clearly delineated clusters, even though there was no supervision involved, other than the teacher features. However, WeCo+KD, which is based on complete weight adaptation during distillation, creates clusters which are spread across the feature space. In summary, the t-SNE visualizations emphasize that WeCoLoRA is an effective few-shot knowledge distillation approach, explaining the superior results on downstream linear probing tasks through the robustness of the learned latent space.

**Attention visualizations.** To visualize the attention, we employ Attention Rollout [1], a technique that enables the visualization of the attention flow throughout ViT. We fuse the attention heads by taking the maximum response (top 10%) and discard pixels with lower values. In Figure 3, we present three test images that are randomly taken from ImageNet [15], which correspond to "Daisy", "Persian Cat" and "Goldfish" classes, respectively. The attention is extracted from student models trained with WeCo+KD and WeCoLoRA. We observe that WeCoLoRA makes the model focus more on discriminative features such as petals, fur, and fish scales. This helps the model based on WeCoLoRA in taking informed decisions.

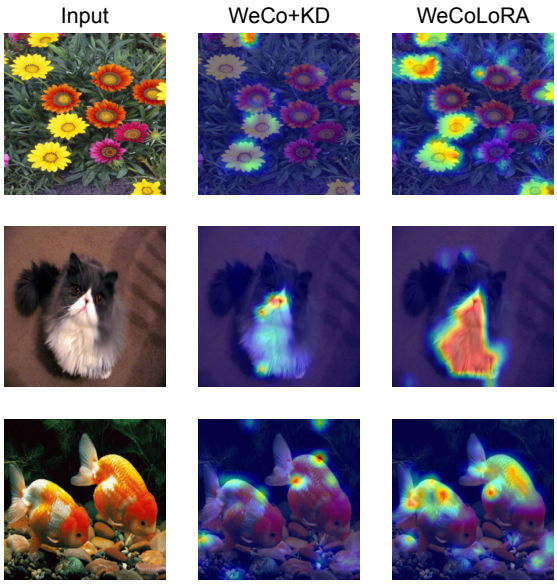


**Fig. 3:** Attention visualizations obtained with Attention Rollout [1] for WeCo+KD and WeCoLoRA, on three test images from ImageNet-1K. The compared students are distilled from the supervised ViT-B teacher [19], with a compression factor of $r = 2$, on 1% of the original training data. The first column displays the original images, the second column displays the attention of the WeCo+KD-based student, and the third column shows the attention of the WeCoLoRA-based student. Best viewed in color.

**Varying the matrix rank.** In Figure 4, we present results obtained by WeCoLoRA on the CIFAR-100 data set, when changing the matrix rank $k \in \{64, 128, 256, 512\}$. The teacher is the supervised ViT-B [19], the compression factor is $r = 2$, and the distillation process is based on 10% of the original training data. We observe that WeCoLoRA obtains stable performance when $k = 128$ and $k = 256$. If $k$ is too small ($k = 64$), the performance drops, showing that the model needs more parameters to recover the information of the skipped teacher layers. However, the performance also drops when $k$ is too large ($k = 512$), but this happens because the model starts to overfit the small data set. Hence, we observe that $k$ is a hyperparameter that can control the bias-variance trade-off of WeCoLoRA. In our experiments, we found that $k = 128$ and $k = 256$ perform generally well in the few-shot distillation scenarios.

**Computational evaluation.** We first compare the running time of the teacher ViT-B and the student based on a compression ratio $r = 2$. As expected, the student is twice as fast as its teacher, evaluating 512 samples in 3 `ms`, while the teacher model requires 6 `ms` for the same mini-batch size. Next, we compare WeCoLoRA and WeCo+KD in terms of the number of trainable parameters. WeCo+KD updates 44M parameters, while WeCOLoRA updates only 10M parameters (when $k = 128$). Because the new weights added by WeCoLoRA are not integrated into the network during training, WeCoLoRA uses 10.6 GFLOPs, while WeCo+KD utilizes 8.5 GFLOPs. During inference, the GFLOPs are the

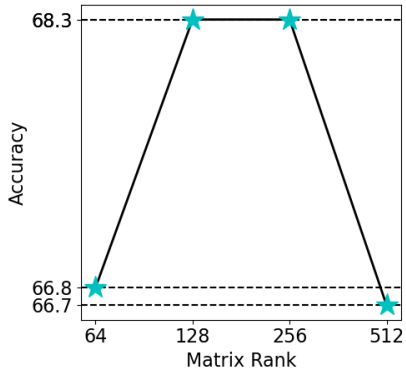


**Fig. 4:** Accuracy rates of WeCoLoRA on the CIFAR-100 data set [34] when varying the matrix rank. The teacher is the supervised ViT-B [19], the compression factor is $r = 2$, and the distillation process uses 10% of the ImageNet-1K training set [15].

same for both WeCoLoRA and WeCo+KD. The reported running times are measured on a GeForce RTX 3090 GPU with 24G of VRAM.

## 5 Limitations

The main limitation of our method is its applicability to vision transformers and architectures that use multiple consecutive blocks with the same configuration, *e.g.* ResNets [26]. This restriction is imposed by our weight copying mechanism. Our ablation results indicate that the weight copying step is very useful in the few-shot distillation scenario, as it significantly boosts performance (see Table 1). Simply removing the weight copying step is not a viable option, since the performance would drastically degrade. To make our framework applicable to any architecture, the weight copying mechanism could be enhanced with adaptor blocks, which would be able to reshape the copied weights to the appropriate size. However, the adaptor blocks need to be tailored for each specific pair of teacher and student models. This will increase the complexity of the hyperparameter tuning stage, which, in the current form, is quite straightforward, *i.e.* aside from typical hyperparameters, such as the learning rate and the mini-batch size, WeCoLoRA only adds the compression ratio $r$ and the rank of the low-rank matrices $k$ as extra hyperparameters.

## 6 Conclusion

In this paper, we proposed a novel few-shot unsupervised feature distillation method that can be used to train vision transformers on a few unlabeled images. Our approach, termed WeCoLoRA, combines weight copying and low-rank adaptation in an efficient and effective training pipeline. We conducted experiments in multiple few-shot scenarios, using both supervised and self-supervised teachers, and considering various model compression factors. The results show

that our method outperforms state-of-the-art competitors [3, 62], as well as ablated versions of our approach. Moreover, we present feature visualizations that clearly indicate that WeCoLoRA produces more robust embeddings, which are able to better disentangle the classes, even on downstream data sets.

In future work, we aim to extend WeCoLoRA to language and audio transformers, which, in principle, should be possible without major updates. Furthermore, we aim to generalize the weight copying mechanism to other architectures, besides transformers, to make WeCoLoRA applicable to a broader variety of models.

## Acknowledgment

The research leading to these results has received funding from the NO Grants 2014-2021, under project ELO-Hyp contract no. 24/2020.

## References

1. Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. In: Proceedings of ACL. pp. 4190–4197 (2020) 12, 13
2. Ba, J., Caruana, R.: Do deep nets really need to be deep? In: Proceedings of NeurIPS. vol. 27 (2014) 3
3. Bai, Y., Wang, Z., Xiao, J., Wei, C., Wang, H., Yuille, A.L., Zhou, Y., Xie, C.: Masked autoencoders enable efficient knowledge distillers. In: Proceedings of CVPR. pp. 24256–24265 (2023) 3, 4, 10, 15, 19, 20
4. Bao, H., Dong, L., Piao, S., Wei, F.: BEiT: BERT Pre-Training of Image Transformers. In: Proceeding of ICLR (2022) 1
5. Bărbălău, A., Cosma, A., Ionescu, R.T., Popescu, M.: Black-Box Ripper: Copying black-box models using generative evolutionary algorithms. In: Proceedings of NeurIPS. vol. 33, pp. 20120–20129 (2020) 4
6. Buciluǎ, C., Caruana, R., Niculescu-Mizil, A.: Model compression. In: Proceedings of KDD. pp. 535–541 (2006) 3
7. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Proceedings of ECCV. pp. 213–229 (2020) 1
8. Chawla, A., Yin, H., Molchanov, P., Alvarez, J.: Data-free knowledge distillation for object detection. In: Proceedings of WACV. pp. 3289–3298 (2021) 4
9. Chen, H., Wang, Y., Xu, C., Yang, Z., Liu, C., Shi, B., Xu, C., Xu, C., Tian, Q.: Data-free learning of student networks. In: Proceedings of ICCV. pp. 3514–3522 (2019) 4
10. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. arXiv preprint arXiv:2102.04306 (2021) 1
11. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of CVPR. pp. 1290–1299 (2022) 1
12. Cheng, G., Han, J., Lu, X.: Remote Sensing Image Scene Classification: Benchmark and State of the Art. Proceedings of the IEEE **105**(10), 1865–1883 (2017) 3, 7, 8, 11, 12, 20, 21, 23

13. Cheng, X., Rao, Z., Chen, Y., Zhang, Q.: Explaining knowledge distillation by quantifying the knowledge. In: Proceedings of CVPR. pp. 12925–12935 (2020) 3
14. Cho, J.H., Hariharan, B.: On the efficacy of knowledge distillation. In: Proceedings of ICCV. pp. 4793–4801 (2019) 3
15. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: Proceedings of CVPR. pp. 248–255 (2009) 3, 7, 9, 10, 11, 12, 14, 20, 21, 22, 23, 24
16. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of NAACL. pp. 4171–4186 (2019) 2
17. Do, K., Le, T.H., Nguyen, D., Nguyen, D., Harikumar, H., Tran, T., Rana, S., Venkatesh, S.: Momentum adversarial distillation: Handling large distribution shifts in data-free knowledge distillation. In: Proceedings of NeurIPS. vol. 35, pp. 10055–10067 (2022) 4
18. Dong, X., Bao, J., Zhang, T., Chen, D., Zhang, W., Yuan, L., Chen, D., Wen, F., Yu, N., Guo, B.: PeCo: Perceptual Codebook for BERT Pre-training of Vision Transformers. In: Proceedings of AAAI. pp. 552–560 (2023) 2
19. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of ICLR (2021) 1, 2, 5, 8, 10, 11, 12, 13, 14, 19, 20, 21, 22, 23, 24
20. Gao, Y., Zhou, M., Metaxas, D.: UTNet: A Hybrid Transformer Architecture for Medical Image Segmentation. In: Proceedings of MICCAI. pp. 61–71 (2021) 1
21. Georgescu, M.I., Fonseca, E., Ionescu, R.T., Lucic, M., Schmid, C., Arnab, A.: Audiovisual masked autoencoders. In: Proceedings of ICCV. pp. 16144–16154 (2023) 2
22. Hao, Z., Guo, J., Jia, D., Han, K., Tang, Y., Zhang, C., Hu, H., Wang, Y.: Learning efficient vision transformers via fine-grained manifold distillation. In: Proceedings of NeurIPS. vol. 35, pp. 9164–9175 (2022) 4
23. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: UNETR: Transformers for 3D Medical Image Segmentation. In: Proceedings of WACV. pp. 1748–1758 (2022) 1
24. He, J., Ding, Y., Zhang, M., Li, D.: Towards efficient network compression via Few-Shot Slimming. Neural Networks **147**, 113–125 (2022) 2, 4
25. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of CVPR. pp. 16000–16009 (2022) 2, 4, 8, 9, 10, 11, 12, 21
26. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: Proceedings of CVPR. pp. 770–778 (2016) 14
27. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015) 3, 4
28. Hu, E.J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: LoRA: Low-Rank Adaptation of Large Language Models. In: Proceedings of ICLR (2022) 3, 4
29. Ji, G., Zhu, Z.: Knowledge distillation in wide neural networks: Risk bound, data efficiency and imperfect teacher. In: Proceedings of NeurIPS. vol. 33, pp. 20823–20833 (2020) 3
30. Justo, J.A., Garrett, J., Langer, D.D., Henriksen, M.B., Ionescu, R.T., Johansen, T.A.: An Open Hyperspectral Dataset with Sea-Land-Cloud Ground-Truth from the HYPSO-1 Satellite. arXiv preprint arXiv:2308.13679 (2023) 2

31. Kelenyi, B., Domsa, V., Tamas, L.: SAM-Net: Self-Attention based Feature Matching with Spatial Transformers and Knowledge Distillation. Expert Systems with Applications **242**, 122804 (2024) 4
32. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in Vision: A Survey. ACM Computing Surveys **54**(10s), 1–41 (2022) 1
33. Kim, S., Baek, J., Park, J., Kim, G., Kim, S.: InstaFormer: Instance-aware image-to-image translation with transformer. In: Proceedings of CVPR. pp. 18321–18331 (2022) 1
34. Krizhevsky, A.: Learning multiple layers of features from tiny images. Tech. rep., University of Toronto (2009) 3, 7, 9, 10, 11, 14, 19, 20, 21, 23
35. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Communications of the ACM **60**(6), 84–90 (2017) 2
36. Lao, S., Song, G., Liu, B., Liu, Y., Yang, Y.: Masked autoencoders are stronger knowledge distillers. In: Proceedings of ICCV. pp. 6384–6393 (2023) 4
37. Lee, S.H., Lee, S., Song, B.C.: Vision transformer for small-size datasets. arXiv preprint arXiv:2112.13492 (2021) 2
38. Li, G., Zheng, H., Liu, D., Wang, C., Su, B., Zheng, C.: SemMAE: Semantic-Guided Masking for Learning Masked Autoencoders. In: Proceedings of NeurIPS. pp. 14290–14302 (2022) 2
39. Li, T., Li, J., Liu, Z., Zhang, C.: Few Sample Knowledge Distillation for Efficient Network Compression. In: Proceedings of CVPR. pp. 14627–14635 (2020) 2, 4
40. Li, Y., Mao, H., Girshick, R., He, K.: Exploring plain vision transformer backbones for object detection. In: Proceedings of ECCV. pp. 280–296 (2022) 1
41. Lim, J.Y., Lim, K.M., Ooi, S.Y., Lee, C.P.: Efficient-PrototypicalNet with self knowledge distillation for few-shot learning. Neurocomputing **459**, 327–337 (2021) 2, 4
42. Lin, H., Han, G., Ma, J., Huang, S., Lin, X., Chang, S.F.: Supervised masked knowledge distillation for few-shot transformers. In: Proceedings of CVPR. pp. 19649–19659 (2023) 2, 4
43. Liu, R., Fusi, N., Mackey, L.: Model compression with generative adversarial networks. In: Proceedings of SGOML (2018) 4
44. Liu, Y., Sangineto, E., Bi, W., Sebe, N., Lepri, B., Nadai, M.: Efficient training of visual transformers with small datasets. In: Proceedings of NeurIPS. pp. 23818–23830 (2021) 2
45. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In: Proceedings of ICCV. pp. 10012–10022 (2021) 1, 2
46. Lopes, R.G., Fenu, S., Starner, T.: Data-free knowledge distillation for deep neural networks. arXiv preprint arXiv:1710.07535 (2017) 4
47. Lopez-Paz, D., Bottou, L., Schölkopf, B., Vapnik, V.: Unifying distillation and privileged information. arXiv preprint arXiv:1511.03643 (2015) 3
48. Luo, S., Tan, Y., Patil, S., Gu, D., von Platen, P., Passos, A., Huang, L., Li, J., Zhao, H.: LCM-LoRA: A Universal Stable-Diffusion Acceleration Module. arXiv preprint arXiv:2311.05556 (2023) 5, 7, 9
49. Madan, N., Ristea, N.C., Nasrollahi, K., Moeslund, T.B., Ionescu, R.T.: CL-MAE: Curriculum-Learned Masked Autoencoders. In: Proceedings of WACV. pp. 2492–2502 (2024) 1
50. Mirzadeh, S.I., Farajtabar, M., Li, A., Levine, N., Matsukawa, A., Ghasemzadeh, H.: Improved knowledge distillation via teacher assistant. In: Proceedings of AAAI. vol. 34, pp. 5191–5198 (2020) 3

51. Nguyen, D., Gupta, S., Do, K., Venkatesh, S.: Black-box few-shot knowledge distillation. In: Proceedings of ECCV. pp. 196–211 (2022) 2, 4
52. Oymak, S., Fabian, Z., Li, M., Soltanolkotabi, M.: Generalization Guarantees for Neural Networks via Harnessing the Low-rank Structure of the Jacobian. arXiv preprint arXiv:1906.05392 (2019) 4
53. Pan, H., Wang, C., Qiu, M., Zhang, Y., Li, Y., Huang, J.: Meta-KD: A Meta Knowledge Distillation Framework for Language Model Compression across Domains. In: Proceedings of ACL. pp. 3026–3036 (2021) 2
54. Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., Tran, D.: Image transformer. In: Proceedings of ICML. pp. 4055–4064. PMLR (2018) 1
55. Rajasegaran, J., Khan, S., Hayat, M., Khan, F.S., Shah, M.: Self-supervised knowledge distillation for few-shot learning. In: Proceedings of BMVC (2021) 2, 4
56. Ristea, N.C., Miron, A.I., Savencu, O., Georgescu, M.I., Verga, N., Khan, F.S., Ionescu, R.T.: CyTran: A cycle-consistent transformer with multi-level consistency for non-contrast to contrast CT translation. Neurocomputing **538**, 126211 (2023) 1
57. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: FitNets: Hints for Thin Deep Nets. arXiv preprint arXiv:1412.6550 (2014) 3
58. Sauer, A., Asaadi, S., Küch, F.: Knowledge Distillation Meets Few-Shot Learning: An Approach for Few-Shot Intent Classification Within and Across Domains. In: Proceedings of NLP4ConvAI. pp. 108–119 (2022) 2
59. Shen, C., Wang, X., Yin, Y., Song, J., Luo, S., Song, M.: Progressive Network Grafting for Few-Shot Knowledge Distillation. In: Proceedings of AAAI. vol. 35, pp. 2541–2549 (2021) 2, 4
60. Tong, Z., Song, Y., Wang, J., Wang, L.: VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In: Proceedings of NeurIPS. vol. 35, pp. 10078–10093 (2022) 2
61. Torbunov, D., Huang, Y., Yu, H., Huang, J., Yoo, S., Lin, M., Viren, B., Ren, Y.: UVCGAN: UNet Vision Transformer cycle-consistent GAN for unpaired image-to-image translation. In: Proceedings of WACV. pp. 702–712 (2023) 1
62. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jegou, H.: Training data-efficient image transformers & distillation through attention. In: Proceedings of ICML. vol. 139, pp. 10347–10357 (2021) 3, 10, 15, 19, 20
63. Urban, G., Geras, K.J., Kahou, S.E., Aslan, O., Wang, S., Caruana, R., Mohamed, A., Philipose, M., Richardson, M.: Do deep convolutional nets really need to be deep and convolutional? In: Proceedings of ICLR (2016) 3
64. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The iNaturalist Species Classification and Detection Dataset. In: Proceedings of CVPR. pp. 8769–8778 (2018) 3, 7, 8, 9, 10, 11, 20, 21, 22
65. Vapnik, V., Vashist, A.: A new learning paradigm: Learning using privileged information. Neural networks **22**(5-6), 544–557 (2009) 3
66. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Proceedings of NIPS. pp. 5998–6008 (2017) 2
67. Wang, R., Chen, D., Wu, Z., Chen, Y., Dai, X., Liu, M., Yuan, L., Jiang, Y.G.: Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. In: Proceedings of CVPR. pp. 6312–6322 (2023) 2
68. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.: ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Clas-

sification and Localization of Common Thorax Diseases. In: Proceedings of CVPR. pp. 3462–3471 (2017) 3, 7, 8, 11, 20, 21, 24
69. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: CvT: Introducing Convolutions to Vision Transformers. In: Proceedings of ICCV. pp. 22–31 (2021) 1, 2
70. Wu, K., Zhang, J., Peng, H., Liu, M., Xiao, B., Fu, J., Yuan, L.: TinyViT: Fast pretraining distillation for small vision transformers. In: Proceedings of ECCV. pp. 68–85 (2022) 4
71. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: SegFormer: Simple and efficient design for semantic segmentation with transformers. In: Proceedings of NeurIPS. pp. 12077–12090 (2021) 1
72. Xu, Y., Xie, L., Gu, X., Chen, X., Chang, H., Zhang, H., Chen, Z., Zhang, X., Tian, Q.: QA-LoRA: Quantization-Aware Low-Rank Adaptation of Large Language Models. In: Proceedings of ICLR (2024) 4
73. Zhang, G., Kang, G., Yang, Y., Wei, Y.: Few-shot segmentation via cycle-consistent transformer. In: Proceedings of NeurIPS. pp. 21984–21996 (2021) 1
74. Zhang, J., Peng, H., Wu, K., Liu, M., Xiao, B., Fu, J., Yuan, L.: MiniViT: Compressing Vision Transformers with Weight Multiplexing. In: Proceedings of CVPR. pp. 12145–12154 (2022) 4, 19, 20
75. Zhang, L., Shen, L., Ding, L., Tao, D., Duan, L.Y.: Fine-tuning global model via data-free knowledge distillation for non-iid federated learning. In: Proceedings of CVPR. pp. 10174–10183 (2022) 4
76. Zhang, M., Wang, D., Gai, S.: Knowledge distillation for model-agnostic meta-learning. In: Proceedings of ECAI. pp. 1355–1362 (2020) 2, 4
77. Zhao, B., Song, R., Liang, J.: Cumulative spatial knowledge distillation for vision transformers. In: Proceedings of ICCV. pp. 6146–6155 (2023) 4
78. Zheng, M., Gao, P., Wang, X., Li, H., Dong, H.: End-to-end object detection with adaptive clustering transformer. In: Proceedings of BMVC (2020) 1
79. Zhou, W., Xu, C., McAuley, J.: BERT Learns to Teach: Knowledge Distillation with Meta Learning. In: Proceedings of ACL. pp. 7037–7049 (May 2022) 2
80. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: Deformable Transformers for End-to-End Object Detection. In: Proceedings of ICLR (2020) 1
81. Zhu, Y., Wichers, N., Lin, C.C., Wang, X., Chen, T., Shu, L., Lu, H., Liu, C., Luo, L., Chen, J., et al.: SiRA: Sparse Mixture of Low Rank Adaptation. arXiv preprint arXiv:2311.09179 (2023) 4
82. Zhu, Z., Hong, J., Zhou, J.: Data-free knowledge distillation for heterogeneous federated learning. In: Proceedings of ICML. pp. 12878–12889 (2021) 4

## 7 Supplementary

In Table 5, we compare our ablated WeCo+KD and WeCoLoRA with three state-of-the-art knowledge distillation techniques namely, DMAE [3], DeiT [62] and MiniViT [74] using the supervised ViT-B [19] teacher. In most of the cases (except on CIFAR-100 [34], with a compression factor of $r = 2$ and the percentage of the training data $\alpha = 1$) our ablated version, WeCo+KD, surpasses all three state-of-the-art methods [3,62,74]. Our method, WeCoLoRA, outperforms all methods, regardless of the data set. Based on the results reported in Table 5, we conclude that the current knowledge distillation methods are not able to cope with a limited training set.

**Table 5:** Accuracy rates (in percentages) obtained on ImageNet-1K [15], iNaturalist19 [64] and CIFAR-100 [34] by three state-of-the-art knowledge distillation methods, DMAE [3], DeiT [62] and MiniViT [74], in comparison with WeCo+KD and WeCoLoRA. The results are obtained using a compression factor of $r = 2$, employing the supervised ViT-B [19] teacher. The parameter $\alpha$ represents the percentage of the original ImageNet-1K training set [15] used during knowledge distillation. The best accuracy on each downstream data set and each value of $\alpha$ is highlighted in bold.

| $\alpha$ | Method | Venue | ImageNet | iNaturalist | CIFAR-100 |
|---|---|---|---|---|---|
| $\alpha = 1$ | DeiT [62] | ICML 2021 | 25.5 | 20.3 | 39.0 |
| | MiniViT [74] | CVPR 2022 | 26.7 | 21.0 | 37.3 |
| | DMAE [3] | CVPR 2023 | 15.4 | 7.9 | 22.2 |
| | WeCo+KD (ours, ablated) | - | 35.5 | 25.4 | 34.8 |
| | WeCoLoRA (ours) | - | **60.5** | **43.4** | **60.6** |
| $\alpha = 10$ | DeiT [62] | ICML 2021 | 57.6 | 37.8 | 61.4 |
| | MiniViT [74] | CVPR 2022 | 64.0 | 40.8 | 64.2 |
| | DMAE [3] | CVPR 2023 | 33.0 | 21.0 | 35.6 |
| | WeCo+KD (ours, ablated) | - | 68.3 | 45.7 | 64.4 |
| | WeCoLoRA (ours) | - | **69.2** | **49.5** | **68.3** |

In Tables 6 and 7, we report results on five downstream tasks, when the students use only 2% and 5% of the ImageNet data during distillation. As noted in the main manuscript, we perform linear probing to demonstrate that our method transfers strong features. We notice that our method, WeCoLoRA, attains higher performance than the WeCo+KD method, especially when the distillation data is scarce. We observe a substantial improvement (of at least 2%) on the ImageNet downstream task, regardless of the reduction ratio or the distillation training size, when the teacher is the supervised ViT-B [19] model. We observe the same trend on the other data sets employed in the evaluation. We further note that the features learned by our distillation method also transfer to out-of-distribution data sets, such as ChestX-ray14 [68]. We consider ChestX-ray14 as out-of-distribution because it contains medical images, while the pre-training data set, ImageNet, contains natural images.

We conclude that the proposed distillation method, WeCoLoRA, is robust and obtains improved performance on multiple downstream tasks, especially when the pre-training data set is small. We also emphasize that our method does not require labeled data, and is able to compress both supervised and self-supervised models.

To better assess the performance trends on various downstream tasks when the number of samples increases from 1% to 10%, we further illustrate the performance levels obtained by WeCoLoRA vs. WeCo+KD on ImageNet-1K [15], iNaturalist [64], NWPU-RESISC45 [12], CIFAR-100 [34] and ChestX-ray14 [68] in Figures 5, 6, 7, 8, and 9, respectively. We observe that WeCoLoRA obtains significantly higher performance than WeCo+KD when there is less data involved in the knowledge distillation process (1% and 2% of the original training set [15]). Moreover, in most of the cases, WeCoLoRA also outperforms WeCo+KD when 10% of the original training set in used during knowledge distillation.

**Table 6:** Results of WeCoLoRA and WeCo+KD in terms of accuracy (in percentages) on ImageNet-1K [15], iNaturalist [64], NWPU-RESISC45 [12] and CIFAR-100 [34], and in terms of mean AUC (in percentages) on ChestX-ray14 [68]. Results are reported for the supervised ViT-B [19] teacher and the self-supervised (SSL) VIT-B [25] teacher. During the distillation procedure, only **2**% of the ImageNet-1K training set [15] is used.

| Teacher | Compression factor $r$ | Distillation method | ImageNet | ChestX-ray14 | iNaturalist | RESISC45 | CIFAR-100 |
|---|---|---|---|---|---|---|---|
| ViT-B [19] (supervised) | 2 | WeCo+KD | 46.9 | 68.3 | 35.1 | 61.6 | 42.0 |
| | | WeCoLoRA | **63.5** | **70.0** | **46.5** | **68.5** | **62.9** |
| | 3 | WeCo+KD | 37.0 | 67.8 | 28.2 | 59.8 | 37.9 |
| | | WeCoLoRA | **39.6** | **68.1** | **29.5** | **61.5** | **38.6** |
| ViT-B [25] (SSL) | 2 | WeCo+KD | 46.7 | 68.6 | 28.3 | **62.9** | 41.5 |
| | | WeCoLoRA | **48.2** | **68.9** | **28.5** | 58.8 | **44.8** |
| | 3 | WeCo+KD | 33.6 | 66.5 | 20.0 | 53.7 | 35.7 |
| | | WeCoLoRA | **35.3** | **67.0** | **22.2** | **56.0** | **36.8** |

**Table 7:** Results of WeCoLoRA and WeCo+KD in terms of accuracy (in percentages) on ImageNet-1K [15], iNaturalist [64], NWPU-RESISC45 [12] and CIFAR-100 [34], and in terms of mean AUC (in percentages) on ChestX-ray14 [68]. Results are reported for the supervised ViT-B [19] teacher and the self-supervised (SSL) VIT-B [25] teacher. During the distillation procedure, only **5**% of the ImageNet-1K training set [15] is used.

| Teacher | Compression factor $r$ | Distillation method | ImageNet | ChestX-ray14 | iNaturalist | RESISC45 | CIFAR-100 |
|---|---|---|---|---|---|---|---|
| ViT-B [19] (supervised) | 2 | WeCo+KD | 65.3 | 69.4 | 45.4 | **73.5** | 60.1 |
| | | WeCoLoRA | **67.3** | **70.0** | **49.0** | 69.9 | **66.6** |
| | 3 | WeCo+KD | 52.2 | 68.4 | 37.0 | **66.7** | 46.9 |
| | | WeCoLoRA | **55.3** | **69.9** | **40.2** | 66.1 | **51.7** |
| ViT-B [25] (SSL) | 2 | WeCo+KD | 51.0 | 69.3 | 29.9 | **61.7** | 47.6 |
| | | WeCoLoRA | **54.0** | **69.5** | **32.9** | 61.5 | **47.7** |
| | 3 | WeCo+KD | 36.1 | 66.6 | 18.7 | 51.3 | 30.5 |
| | | WeCoLoRA | **37.4** | **66.5** | **22.0** | **58.6** | **38.6** |

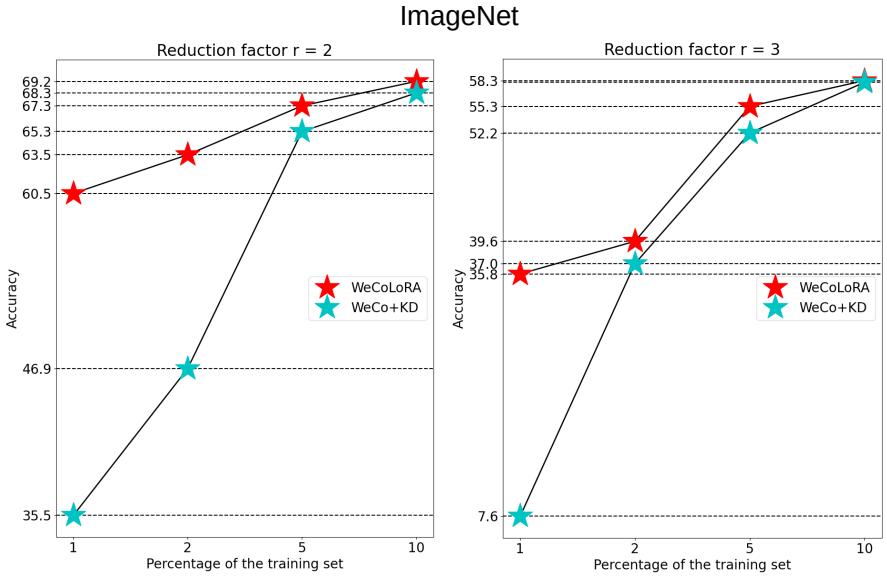


**Fig. 5:** Accuracy rates obtained by WeCoLoRA and WeCo+KD on the ImageNet-1K [15] downstream task. Results are reported for the supervised ViT-B [19] teacher. The horizontal axis corresponds to the percentage of the original training set [15] used during knowledge distillation. Best viewed in color.

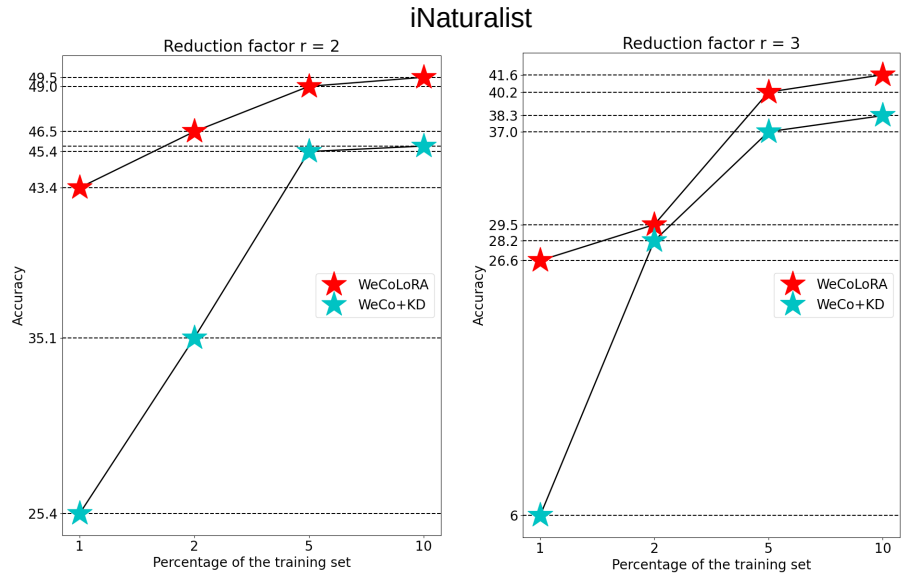


**Fig. 6:** Accuracy rates obtained by WeCoLoRA and WeCo+KD on the iNaturalist [64] downstream task. Results are reported for the supervised ViT-B [19] teacher. The horizontal axis corresponds to the percentage of the original training set [15] used during knowledge distillation. Best viewed in color.

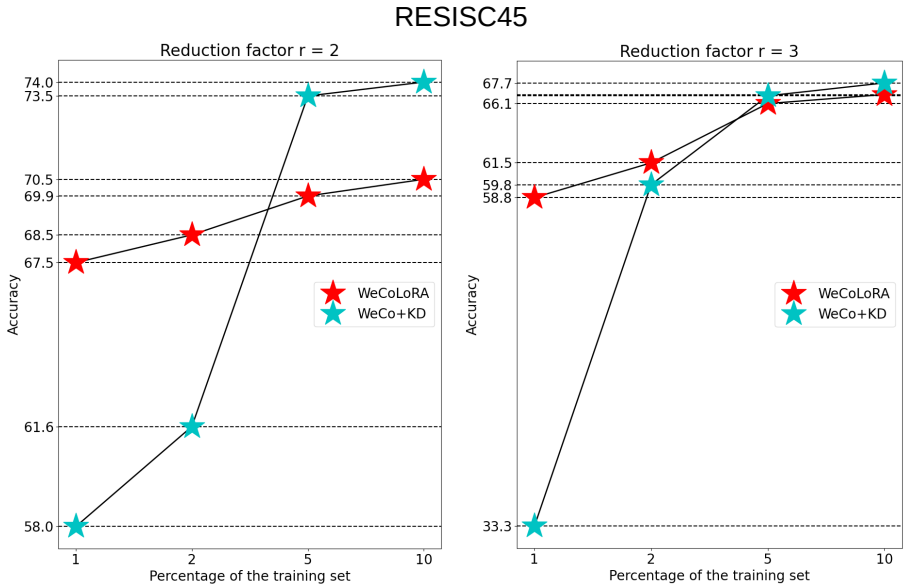


**Fig. 7:** Accuracy rates obtained by WeCoLoRA and WeCo+KD on the NWPU-RESISC45 [12] downstream task. Results are reported for the supervised ViT-B [19] teacher. The horizontal axis corresponds to the percentage of the original training set [15] used during knowledge distillation. Best viewed in color.

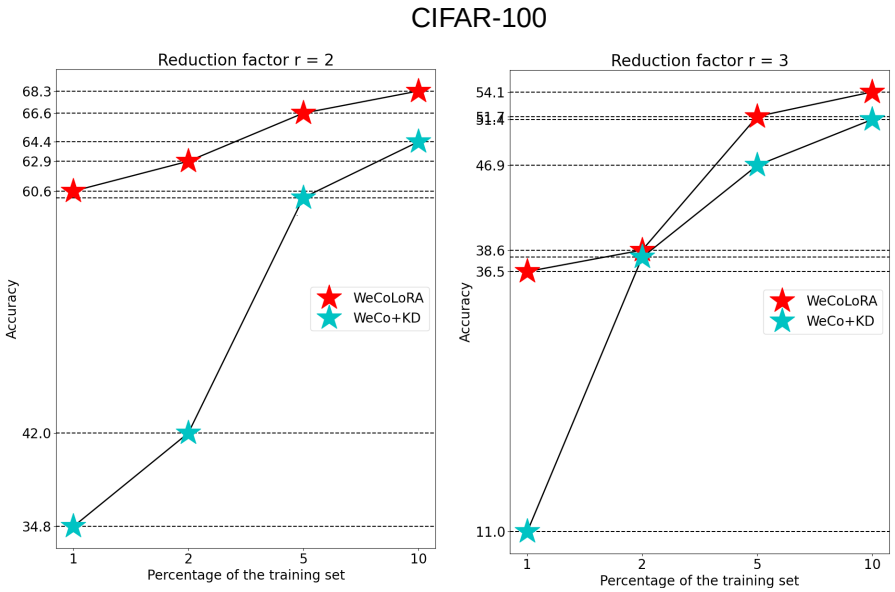


**Fig. 8:** Accuracy rates obtained by WeCoLoRA and WeCo+KD on the CIFAR-100 [34] downstream task. Results are reported for the supervised ViT-B [19] teacher. The horizontal axis corresponds to the percentage of the original training set [15] used during knowledge distillation. Best viewed in color.

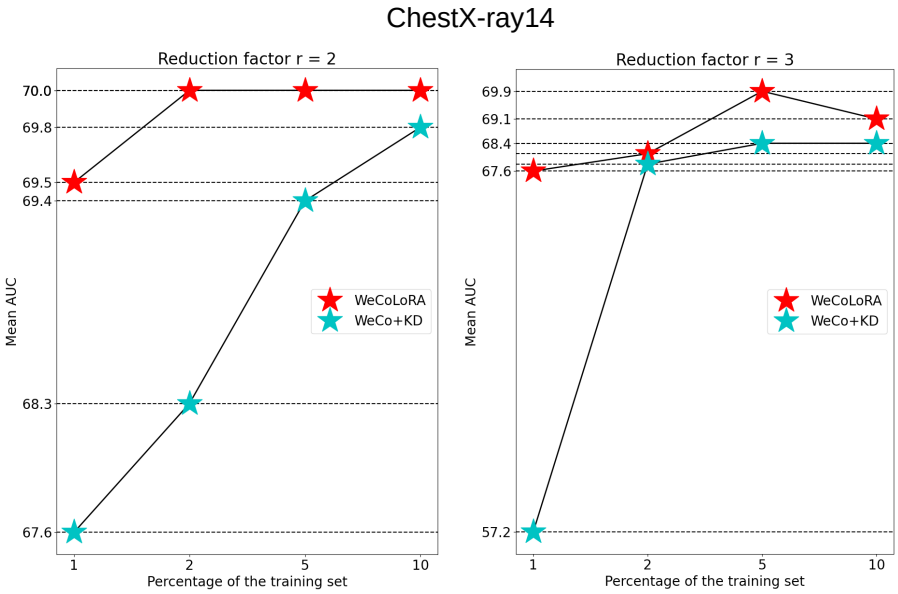


**Fig. 9:** Mean AUC scores (in percentages) obtained by WeCoLoRA and WeCo+KD on the ChestX-ray14 [68] downstream task. Results are reported for the supervised ViT-B [19] teacher. The horizontal axis corresponds to the percentage of the original training set [15] used during knowledge distillation. Best viewed in color.