UNIC: Universal Classification Models via Multi-teacher Distillation

Mert Bülent Sarıyıldız, Philippe Weinzaepfel, Thomas Lucas, Diane Larlus, and Yannis Kalantidis

NAVER LABS Europe
https://europe.naverlabs.com/unic

Abstract. Pretrained models have become a commodity and offer strong results on a broad range of tasks. In this work, we focus on classification and seek to learn a unique encoder able to take from several complementary pretrained models. We aim at even *stronger generalization* across a variety of classification tasks. We propose to learn such an encoder via multi-teacher distillation. We first thoroughly analyze standard distillation when driven by multiple strong teachers with complementary strengths. Guided by this analysis, we gradually propose improvements to the basic distillation setup. Among those, we enrich the architecture of the encoder with a ladder of expendable projectors, which increases the impact of intermediate features during distillation, and we introduce teacher dropping, a regularization mechanism that better balances the teachers' influence. Our final distillation strategy leads to student models of the same capacity as any of the teachers, while retaining or improving upon the performance of the best teacher for each task.

Keywords: Multi-Teacher Distillation · Classification · Generalization

1 Introduction

Recent years have witnessed the rise of many pretrained models [8, 61, 81]. They often share the same architecture and sometimes even the same training data. They generalize to a broad range of tasks, but may particularly excel at specific visual recognition scenarios depending on the selected learning strategy. Self-supervised learning models [8–10] shine in transfer learning, *i.e.* generalization to novel classes, while models trained with masked modeling techniques [18, 81] are often better suited to patch-level tasks. Meanwhile, supervised learning [14, 29] is still best for specific classification tasks when labeled data is available during pretraining.

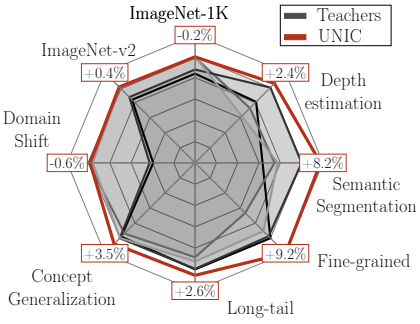


Fig. 1: Relative gains using our UNIC encoder distilled from four teachers (DINO, DeiT-III, iBOT, dBOT-ft), over the respective best teacher for each task using a *single encoder* and no task-specific parameters.

In this paper, our goal is to learn *a universal encoder* capable of strong generalization across a broad spectrum of classification tasks. More specifically, besides ImageNet classification [52] – the dataset on which our teachers are trained and our students are distilled – we are further interested in the classification of novel classes, on new domains, as well as dense prediction tasks such as semantic segmentation or depth estimation. Our goal is to learn *a single encoder* that can be directly applied to all these tasks, out-of-the-box, without the need for any task-specific parameters besides a linear classifier per classification task.

Our approach uses multi-teacher distillation, drawing on the strengths of various specialized teachers to train an encoder that seeks to match or surpass the best teacher in each task. We conduct a comprehensive analysis of the distillation process from multiple teachers, evaluating our models on various tasks, including image-level classification on ImageNet-1K and 15 more transfer datasets, as well as patch-level classification tasks such as semantic segmentation and depth estimation. We leverage our findings to gradually devise a method that shows improved generalization across multiple tasks and axes. We modify the input of expandable projectors [9, 10, 54] (building what we call a *ladder of projectors*) so that they also act as information highways that propagate signal from intermediate layers to the distillation loss in a more direct manner. We analyze learning dynamics across teachers and further propose *teacher dropping*, an effective strategy for balancing the teachers' influence in multi-teacher distillation, resulting in significant gains for the tasks at which our distilled models were otherwise underperforming.

With all of our improvements added to the basic multi-teacher distillation setup, we are able to train models that exhibit strong generalization across a wide range of classification tasks on the image and patch levels, either retaining or improving the performance of the best teacher. As an example, we show in Fig. 1 that by distilling from four strong ViT-Base models trained on ImageNet (*i.e.* DINO [8], DeiT-III [62], iBOT [81], and dBOT-ft [31]) we are able to train a *universal encoder* excelling at all considered tasks. In our experimental study, we show that our findings further extend to the case of larger teachers like DINOv2 [39] and MetaCLIP [70] trained on arbitrary datasets. Finally, we study the way the distilled encoders utilize their weights: first, by quantifying performance drops after weights pruning, and second after reducing the dimension of the output feature space using PCA. These experiments show that distilled models have lower redundancy in both their weights and their features.

Contributions. To summarize, we conduct a thorough analysis of multi-teacher distillation for ViT encoders and use our findings to improve the distillation process and generalization power of the student. Among other simple but crucial modifications, we introduce improvements like ladder of projectors and teacher dropping regularization that enable us to learn models which retain or improve the performance of the best teachers across many diverse tasks. We refer to such models as **Uni**versal **C**lassification models or **UNIC**. We finally perform extensive evaluations along multiple axes of generalization and study the ways the resulting models make use of their weights and feature space.

2 Related Work

Knowledge distillation (KD) was initially introduced as a model compression technique [7], where the goal is to train a smaller student model from the output of a teacher model [23]. While early work focused on predicting the final outputs of a classification model, the idea was rapidly extended to other forms of distillation, such as distilling intermediate representations [1,21,22,49,73,75,79]. These methods perform well but require careful layer selection and loss balancing [21]. In our work, instead of matching layer-wise representations between the student and teacher architectures, we add shortcut connections from intermediate layers of the student to the loss of each teacher.

Multi-teacher knowledge distillation. KD can naturally be extended to an ensemble of teachers so that student can benefit from their potential complementarity. While the final outputs of teachers trained for the same task can simply be averaged [3,15,23,75], multi-teacher distillation with teachers trained for different tasks is more challenging. UDON [76] first trains domain-specialist teachers which are subsequently distilled in a student model using adaptive data sampling for balancing the different domains. In [60], contrastive learning is used for ensemble distillation while [56] proposes a framework tailored for teachers trained with masked image modeling and contrastive learning. But such approaches are not straightforward to extend to teachers learned differently. Similarly, [71] combines self-supervised teachers from arbitrary heterogeneous pretext tasks. [13, 16, 51] focus on jointly utilizing pseudo- and true labels for multi-teacher distillation. Roth *et al*. [51] formulate multi-teacher distillation as continual learning and further propose a novel method for data partitioning based on confidence. Here we develop a more generic method for combining teachers, that is not limited to certain types of teachers or losses, and, unlike [30, 51], does not require labeled data, nor classifiers associated with each teacher for obtaining pseudo-labels.

Loss balancing is shown to be crucial in multi-task learning [11,24,26,78]. Similar strategies to automatically balance losses have also been proposed for multi-teacher distillation [15,32]. In [24], adaptive loss weights inversely proportional to the average of each loss are introduced, while [32] learns instance-level teacher importance weights using ground-truth labels. In [15], the random selection of one teacher per mini-batch is shown to help. Our experiments show that our proposed generalized teacher dropping strategy leads to better models compared to [15,24].

Distilling from a "foundation model" like CLIP [43] or DINOv2 [39] is an effective approach for tasks with limited training data [36,42,67]. Distilling from *multiple* foundation models allows for more versatile students. Recent works like AM-RADIO [46], SAM-CLIP [65], and Open Vocabulary SAM [77] combine the semantics captured by CLIP with the localization capabilities of models like DINOv2 [39] or SAM [27]. AM-RADIO [46] builds on the same base setup as our study, but employs no loss balancing. Another difference comes from the fact that their student encoder is only a part of the final model: AM-RADIO requires the teacher-specific projectors learned during distillation to also be used at test

time, effectively increasing the parameters of the encoder with task-specific ones. Instead, our method performs well on multiple classification tasks *out-of-the-box*, without any additional parameters.

Combining models beyond distillation. Other creative ways to combine multiple pretrained models have been proposed. Works like [37,44,45,59,68] explore different weight averaging strategies. They typically only combine models that differ by their hyper-parameter configuration. Aiming at generalization, [72] merges multiple ViTs, each specialized to a classification task, into a single encoder that solves all classification tasks jointly, via a gating network with multiple functions. Instead, our students are distilled from scratch, have a simple ViT architecture and tackle diverse classification tasks with simple linear probing.

Expendable projectors are extra modules that act as buffers between the final encoder output and the space where the loss is computed. They have been successfully used for both self-supervised [9,10] and supervised learning [54,66]. We extend this idea and add projectors during training to intermediate layers as well. Roth *et al*. [50] use several such projectors of varied dimensionality for metric learning, but do not use features from intermediate layers. Moreover, we use a specific set of projectors per teacher, similar to [3,46]. This way, projectors become *loss-specific*, *i.e*. they contribute to the loss for only one of the teachers.

3 Analyzing and improving multi-teacher distillation

In this section we first present the multi-teacher distillation setup we use as a basis for our analysis (Sec. 3.1) and a summary of our evaluation protocol (Sec. 3.2). We then delve into challenges around multi-teacher distillation of ViT encoders (Sec. 3.3), and offer improvements to the basic setup to overcome them, like enhanced expendable teacher-specific projectors heads (Sec. 3.4) and strategies to more equally learn from all teachers (Sec. 3.5).

3.1 A basic distillation setup

Our task is to distil M *teacher* models $\mathcal{T} = \{\mathcal{T}_1, \dots, \mathcal{T}_M\}$ into a student model $\mathcal{S}$. An overview is shown in Figure 2. Each teacher $t \in \mathcal{T}$ is a ViT [14] encoder that maps an image $\boldsymbol{x}$ to a set of d-dimensional feature vectors $\boldsymbol{y}_{t,i} = f_t(\boldsymbol{x}; i)$ for token i, which can either be one of the $H \times W$ patch tokens from $\mathcal{P}$ or the global CLS token c. We aim at learning the parameters f_s of the student $\mathcal{S}$, such that the output representations $\boldsymbol{z}_i = f_s(\boldsymbol{x}; i)$ excel at all the tasks that any of the teachers also shines at.

We append a *projector head* h_t per teacher to the student encoder's output which transforms each token into a teacher-specific representation $h_t(\boldsymbol{z}_i)$. The loss for each teacher is then computed on $h_t(\boldsymbol{z}_i)$, the output of the corresponding projector head. We consider these projector heads as *expendable*, *i.e*. they are removed after distillation and are not part of the student encoder. Their goal is to assist the learning process, taking inspiration from similar expendable projectors used in self-supervised [9] and supervised [54,66] representation learning. We set

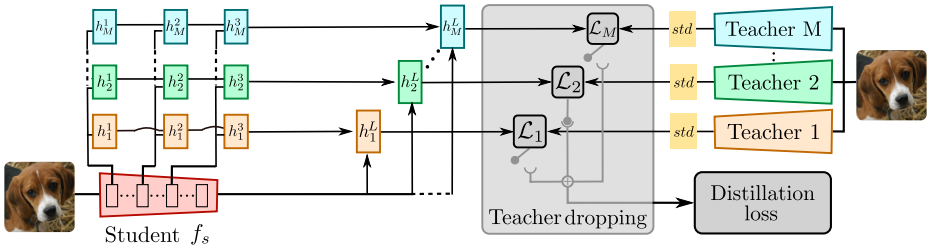


Fig. 2: Overview of our multi-teacher distillation setup. The same input image is fed to each teacher and to student. We employ feature standardization at the output of all teachers (Sec. 3.3), *a ladder of expandable projectors* attached to student (Sec. 3.4) and *teacher dropping regularization* to balance teachers (Sec. 3.5). The latter enables us to adaptively select a subset of teachers to contribute to the loss simply using loss magnitudes. We use dedicated projectors for the CLS and patch tokens (Sec. 3.3).

projector heads to be Multi-Layer Perceptrons (MLPs) with two linear layers, GeLU non-linearity and hidden dimension of $d_h = 4d$, where d is the feature dimension; we analyze projectors further in the next sections.

We use two common distillation losses: cosine and smooth-ℓ_1 (see supplementary material for details); the loss for token i from teacher t is given by:

$$\mathcal{L}_t(\boldsymbol{x}; i) = 0.5 \times \big(\mathcal{L}^{cos}(h_t(\boldsymbol{z}_i), \boldsymbol{y}_{t,i}) + \mathcal{L}^{s\ell_1}(h_t(\boldsymbol{z}_i), \boldsymbol{y}_{t,i})\big). \tag{1}$$

This loss is computed separately for the CLS and each of the patch tokens $\mathcal{P}$. To get the final loss, we sum losses from all teachers similar to [75], as well as over the CLS token c and the tokens of all patches:

$$\mathcal{L}(\boldsymbol{x}) = \sum_{t \in \mathcal{T}} \big(\mathcal{L}_t(\boldsymbol{x}; c) + \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \mathcal{L}_t(\boldsymbol{x}; p)\big), \tag{2}$$

where $|\mathcal{P}|$ is the number of patch tokens.

3.2 Protocol summary

We first present a summary of the experimental protocol we use for the analysis in this section. Further details are presented in the supplementary material.

Datasets and backbones. To better isolate the effects of different distillation components, we use the same training data and architectures for all teachers and students, *i.e.* the ImageNet-1K dataset [52] and ViT-Base [14], respectively. During distillation, we discard the labels of ImageNet and only use the images; no supervised loss is combined with the distillation losses presented above.

Teachers. We consider models learned using *self-supervised learning* (SSL), like DINO [8] or iBOT [81], and *supervised models* like DeiT-III [62] or fine-tuned dBoT [31], optimized for the classification task of ImageNet-1K. The former have proven extremely effective for generalization whereas the latter achieve state-of-the-art accuracy on the ImageNet-1K task. In this section, we present

Table 1: Component analysis for distillation from two teachers. We report: image classification on 1) ImageNet-1K (IN-val) and 2) 15 transfer learning datasets (averaged), 3) semantic segmentation on ADE-20K, and 4) depth estimation on NYUd. Column legend: std: feature standardization, DP: dedicated projector heads for CLS/patch tokens, LP: ladder of projectors and *tdrop*: teacher dropping regularization.

	Model	std	DP	LP	*tdrop*	IN-val top-1 (↑)	Transfer top-1 (↑)	Segmentation mIoU (↑)	Depth RMSE (↓)
	Teacher models								
1.	DINO					77.7	72.4	30.4	0.570
2.	DeiT-III					83.6	68.5	32.3	0.589
3.	*best teacher*					83.6	72.4	32.3	0.570
	Multi-teacher distillation (DINO & DeiT-III teachers)								
4.	basic setup					78.7	73.1	33.9	0.560
5.	**UNIC**	✓				81.4	73.8	36.1	0.558
6.		✓	✓			82.2	74.1	36.9	0.551
7.		✓	✓	✓		82.7	74.2	37.4	0.546
8.		✓	✓	✓	✓	83.2	73.5	37.3	0.547

our analysis for $M = 2$ teachers, specifically DINO and Deit-III. More teachers and combinations are explored in Sec. 4 and in the supplementary material.

Tasks. We measure performance on many tasks, divided along the following axes: 1) Top-1 accuracy on the *training set classes* on the ImageNet-1K validation set [52] (IN-val); 2) *Transfer learning* performance on unseen classes; we report top-1 accuracy averaged over 15 diverse image classification datasets;[1] *Dense prediction* performance on 3) semantic segmentation and 4) depth estimation; we report mIoU on ADE-20k [80] and RMSE on NYUD [57], measured using a protocol that is essentially dense classification, *i.e.* using linear probes as in [39]. We learn linear probes for all tasks directly over encoder outputs $\boldsymbol{z}$.

3.3 Analyzing multi-teacher distillation of ViT tokens

In this section we analyze and revisit different aspects of distillation that are specific to ViT encoders, *e.g.* the use of CLS and patch tokens. The former is normally fed as input to image-level classifiers while patch tokens are important for dense prediction. In this section we study their statistics and explore how this affects design choices of the distillation setup. The top part of Tab. 1 compares the accuracy of the self-supervised DINO and supervised DeiT-III on the different evaluation axes. They show complementary strengths, *i.e.* they respectively perform well on transfer learning and the ImageNet-1K validation set (IN-val).

Equalizing feature statistics across tokens and teachers. We start by analyzing the statistics of features extracted from the CLS and patch tokens of both teachers and show that this should be taken into account for multi-teacher

[1] The 15 datasets are: 5 ImageNet-CoG levels [55] tailored for concept generalization, 8 small-scale fine-grained datasets (Aircraft, Cars196, DTD, EuroSAT, Flowers, Pets, Food101, SUN397) and two long-tail datasets (iNaturalist-2018 and 2019).

distillation. We calculate such statistics and notice a number of discrepancies in their first and second moment values, both between CLS and patch tokens of a given teacher as well as across teachers. The norm and standard deviation for the CLS token features of DINO, for example, are double the ones for patch tokens of the same model, while the same statistics also differ across DeiT-III and DINO tokens (see supplementary material for more details).

To explore whether such statistical inconsistencies across features affect distillation, we add feature standardization on each teacher output, *i.e.* we normalize teacher features to zero mean and unit variance before computing the loss, which was shown to be useful in [21]. This not only equalizes any differences between CLS and patch tokens but also for tokens across teachers. For convenience and generality, we propose to learn such normalization statistics on-the-fly during distillation, using an exponential moving average. From Tab. 1 we see that the performance of models learned via distillation is consistently higher using feature standardization for both image- and patch-level tasks (rows 4 vs. 5).

▶ *Feature standardization improves multi-teacher distillation*

Projector heads for CLS and patch tokens. Beside statistical differences, the CLS and patch token are also conceptually different: CLS is a global token expected to encode image-level semantics whereas the patch tokens encode local information. To better capture these specifics from CLS and patch tokens, we experiment with dedicated teacher-specific projector heads for each type of tokens. This comes at no added cost in practice, since we discard the projectors after distillation. We discuss expendable projectors further in Sec. 3.4. Comparing rows 5 and 6 in Tab. 1 we see that specializing the teacher-specific projector heads to either CLS or patch tokens leads to further gains.

▶ *Dedicated projectors for CLS/patches improve distillation performance*

Classification on ImageNet and novel classes. Results in Tab. 1 show that models learned via multi-teacher distillation lack in terms of ImageNet-1K performance compared to highly optimized models for that specific task, such as DeiT-III (82.2 vs. 83.6). One may suggest that this is due to the fact that we do not use labels during distillation. To test that, we also performed distillation using *only* the DeiT-III model as a teacher. In that case we were able to reach a top-1 accuracy of 83.1% on ImageNet. This is much higher than the 82.2% we get distilling jointly from multiple teachers and we therefore see that there is still space for improvement during distillation itself.

From Tab. 1 we also see that models learned via multi-teacher distillation greatly outperform DINO on transfer learning and classification of novel classes. This is also true for the recent iBOT [81] model, which also achieves state-of-the-art top-1 accuracy, *i.e.* 72.4% on average for transfer learning on our setup.

▶ *Multi-teacher distillation significantly improves generalization*

Multi-teacher distillation for dense prediction. To assess the discriminative power of patch tokens individually, we consider two dense prediction tasks, semantic segmentation and depth prediction, after linear probing. Tab. 1 shows that even the basic multi-teacher distillation setup improves over the best teacher

(row 4). More importantly, performance increases even further (row 6) using standardization and dedicated projectors for the CLS and patch tokens. The student encoder achieves *+4.6% higher mIoU* than the best teacher for segmentation. This result is even more impressive when compared to the performance of models that are targeting improved dense prediction. Our models, which are distilled from teachers trained with supervised and contrastive learning achieve dense prediction performance comparable to models known to excel at dense tasks, *i.e.* models trained via masked patch prediction like iBOT [81]: iBOT achieves 36.6% mIoU on ADE-20K, while our student reaches 36.9%.

► *Multi-teacher distillation improves the discriminative power of patch tokens*

Retaining complementary teacher strengths. From the results in Tab. 1, we see that models learned with our multi-teacher distillation setup and simple modifications like feature standardization and dedicated projectors for CLS/patch tokens are starting to show strong generalization performance on a number of axes. We will use models distilled under this setup as the basis for the rest of our study. Such models seem to retain the complementary strengths of their teachers: They already outperform the best teacher on transfer learning and dense prediction tasks, while also enjoying decent performance on the ImageNet task.

► *Learning from multiple teachers can combine their strengths*

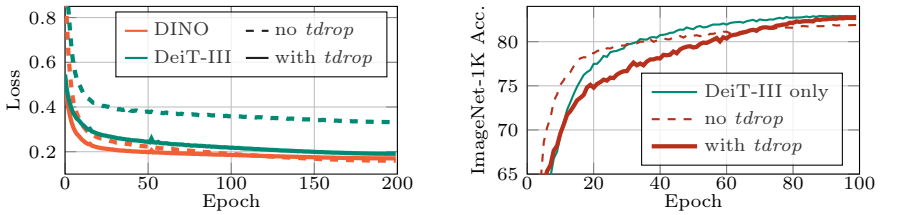
As we discuss above, there is however still room for improvement; we ideally want models to match or outperform the best teacher on all tasks. In the next sections, we analyze different aspects of our distillation setup and introduce further improvements towards that end.

3.4 A ladder of projectors for distillation

The basic setup above uses expendable projector heads as a way of injecting teacher-specific parameters during distillation.[2] Such modules are appended at the end of the encoders and act as small "buffers" between the encoder output and the feature space considered by the loss. In this section, we propose to use more of these expendable modules in a complementary way: as *information highways* that propagate information from intermediate layers to the loss in a more direct manner. Intermediate layers have been used to improve distillation [17, 32, 75], typically by adding extra losses on top of those layers. However, this leads to a more challenging optimization. Besides, hyper-parameter tuning with many added losses is combinatorial, and it becomes cumbersome. These issues are far more prominent in the case of multiple teachers.

Instead of adding losses on intermediate representations, we propose to augment the existing expendable teacher-specific projector head to receive inputs from intermediate layers and append modules that connect all intermediate layer tokens directly to the teacher-specific projector head before the loss. We refer to such augmented projectors as a *ladder of projectors*. This architecture bares similarities to the adaptor architecture that is typically used for adapting a model

[2] Projector heads are discarded after distillation and linear probes are learned over the encoder outputs $\boldsymbol{z}$.



(a) Loss for UNIC with/without *tdrop* **(b)** Performance vs. distilling only DeiT-III

Fig. 3: Analyzing teacher dropping regularization (*tdrop*). **(a)** Loss for each of the two teachers during multi-teacher distillation, with and without *tdrop*. **(b)** ImageNet-1K top-1 accuracy when distilling from DINO & DeiT-III together, versus distilling only from DeiT-III, *i.e.* the teacher that excels at this task.

to a new task [74]. In our case, however, the adaptor-like modules we append during distillation are *expendable*.

Specifically, we attach MLP projectors to intermediate layers and augment the input of the teacher-specific projectors h_t that until now operated only on the last layer of the student encoder. Let $\boldsymbol{z}^l$ denote the l-th layer output of the student encoder for $l = 1, \ldots, L$. The head for the ladder of projectors becomes:

$$h_t^{LP}(\{\boldsymbol{z}^l : l \in L\}) = \sum_{l=1}^{L} h_t^l(\boldsymbol{z}^l), \tag{3}$$

where h_t^l denotes the MLP projector head attached after layer $l \in L$. The architecture of h_t^l is identical to h_t, however, since we are adding multiple such projector heads, we significantly reduce the hidden dimension d_h^l and set $d_h^l = d$ when $l < L$. We explore architecture choices in the supplementary material.

From Tab. 1 we see that this *ladder of projectors* improves performance overall (row 8), especially for dense prediction. It seems that the dense connections lead to better prime patch tokens. Gains are also significant for supervised classification: ImageNet-1K accuracy is increased by +0.5%.

▶ *A ladder of projectors leads to improvements for both CLS and patch tokens*

3.5 Learning all teachers equally well

The basic setup assumes that the final goal is for the distilled encoder to represent each teacher equally well. When distillation uses feature standardization across all teachers and simple losses like cosine and smooth-ℓ_1, there exists a straightforward way to compare how much each of the different teachers is learned: One may simply compare the *magnitudes* of the losses, that indicate how well we are approximating the feature space of each teacher.

Fig. 3a displays the loss curves for multi-teacher distillation for UNIC models, using the setup presented in Sec. 3.3 (dashed lines). We see that the DINO teacher seems to be learned faster and better than DeiT-III.

▶ *Teachers do not equally contribute without further intervention*

It therefore comes as no surprise that our student lacks performance on ImageNet-1K, *i.e.* the task that DeiT-III excels at. But what if DINO was not even part of the distillation process? In Fig. 3b we show how ImageNet-1K accuracy changes during distillation using DINO & DeiT-III as teachers, and for the case of distilling *only* from DeiT-III. We see that our model learns faster using multiple teachers but converges to a lower accuracy: The student seems to exploit features from the additional teacher to ramp up performance faster, but fails to reach the accuracy of distilling DeiT-III alone (83.1%).

Fig. 3 suggests that some form of loss balancing could be beneficial. Loss balancing is common in multi-task settings: In most cases it is done *manually* by adding hyperparameters that control each loss. Such an approach is however cumbersome for many teachers and losses like our case, something also discussed in [46]. It is important to avoid the combinatorial nature of manual tuning. Another way, would be to use some of the existing methods for loss balancing that are proposed for multi-task learning, *e.g.* methods like Adaloss [24]. We argue that the case of multi-teacher distillation over standardized features and simple regression losses is much simpler than multi-task learning when it comes to balancing the losses: The magnitudes of the losses are comparable and can be used for balancing and pacing the distillation process.

Teacher dropping regularization. We introduce a simple scheme for loss balancing that we name *teacher dropping*. Instead of designing some soft loss weighing algorithm, we take inspiration from methods like randomized dropout [58] and path dropping [25], and propose to "drop", *i.e.* zero-out the loss, for a subset of the teachers. Dropping teachers at random is however something that would not encourage loss equalization across teachers. Instead, we propose to directly use absolute magnitudes of the losses when selecting which teachers to drop, *i.e.* keeping the teacher whose loss magnitude is maximal and dropping any other teacher with some probability. This bares conceptual similarities to adaptive dropout [4], but our method is *non-parametric*, and simply exploits the fact that feature space losses on constrained representations are comparable.

We perform loss-based teacher dropping at the image level. At each iteration and for every image, we define a binary coefficient $\alpha_t = \{0, 1\}$ for each teacher t that is multiplied with the corresponding loss $\mathcal{L}_t$. This determines whether teacher t would be dropped or not for that image with probability p. To make sure there is always some signal to learn from, we choose to never drop the teacher with the maximum magnitude loss, *i.e.* the teacher that the current model approximates least well. All other teachers could be dropped with probability p. Specifically and for each image, the coefficient for teacher $t \in \mathcal{T}$ is given by:

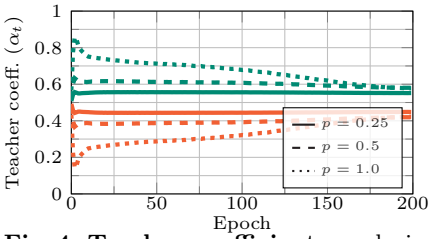


Fig. 4: Teacher coefficients α_t during distillation from DeiT and DINO.

$$\alpha_t = \begin{cases} 1 & \text{if } \mathcal{L}_t = \max_t \mathcal{L}_t, \\ (1-\delta) & \text{if } \mathcal{L}_t \neq \max_i \mathcal{L}_i, \text{ with } \delta \sim \text{Bernoulli}(p) \ . \end{cases} \tag{4}$$

In all cases, the teacher that is least well approximated in the current iteration will always be used. We also experimented with patch-level teacher dropping but found no noticeable gains (see supplementary material).

Effect of teacher dropping during distillation. We study the impact of teacher dropping during distillation in Fig. 3a: teacher dropping makes the loss magnitudes of the teachers much more similar as training progresses (solid lines). In Fig. 4 we plot how the teacher coefficients α_t vary during distillation; teacher utilization becomes more balanced and stabilizes after some epochs.

▶ *Teachers are distilled equally well with teacher dropping regularization*

How does teacher dropping affect performance? We compared teacher dropping regularization to manually balancing the teacher losses, random dropping [15], as well as to the recent Adaloss [24] loss balancing method. Starting from results in row 6 in Tab. 1, we found that none of these strategies is able to noticeably improve, let alone outperform results with teacher dropping (row 8). Specifically, Adaloss achieves 80.1/73.6/34.3/0.565 on the four tasks, respectively (see supplementary material for details). Besides performance, we believe the effectiveness and simplicity of the proposed teacher dropping is unparalleled.

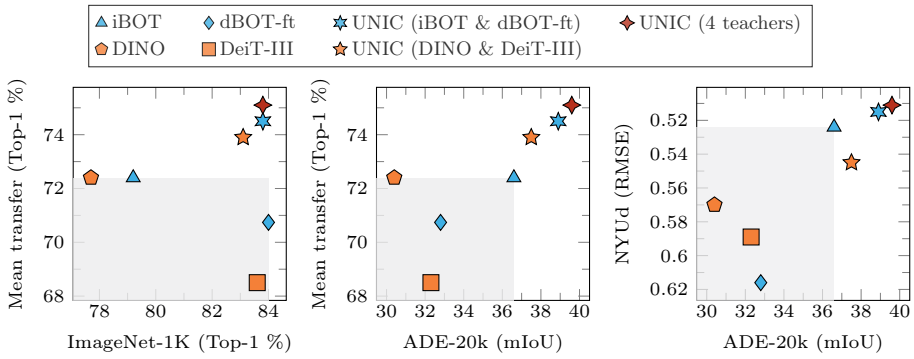
We studied the impact of the teacher dropping probability p and found performances to be stable for different values. Yet, a higher p favours ImageNet performance, with a slight decrease on tasks where the student already outperforms the best teacher (see supplementary material).

From Tab. 1 (row 8) we see that teacher dropping boosts performance for ImageNet-1K, *i.e.* improves distillation on the task where our distilled models were lacking the most. When combining teacher dropping with a ladder of projectors, we are able to achieve 83.2%, our top performance on that task. This performance is only 0.4% lower than the highly optimized DeiT-III (row 3). What is more, we have also closed the observed gap between multi-teacher distillation and specialized distillation using DeiT-III alone. Teacher dropping significantly contributes to that end, *i.e.* increasing performance by 0.5% over our best model with ladder of projectors (rows 7 vs. 8).

▶ *Teacher dropping regularization is a simple and effective way to balance teachers, specifically designed for multi-teacher distillation*

3.6 Towards universal classification models

Multi-teacher distillation using a ladder of projectors and teacher dropping regularization enables us to reach ImageNet classification performance comparable to the highly optimized DeiT-III, while simultaneously outperforming the best teacher on transfer learning performance on 15 datasets with mostly novel classes including long-tail ones, as well as on patch-level classification tasks like semantic segmentation and depth estimation. We contend this evidence demonstrates that our distilled models operate as more *universal* classification models. We will refer to models learned with our enhanced multi-teacher distillation setup as **UNIC** models (which stands for UNIversal Classification, pronounced "*unique*").



(a) Transfer vs. ImageNet **(b)** Transfer vs. Segmentation **(c)** Depth vs. Segmentation

Fig. 5: Performance of different UNIC encoders on different pairs of tasks. We report performance for UNIC encoders distilled from DINO & DeiT-III, iBOT & dBOT-ft and all four teachers. We show results on ImageNet-1K **(a)**, over 15 transfer learning tasks **(a, b)**, semantic segmentation **(b, c)** and depth estimation **(c)**.

4 Experimental study

Teachers. We report results distilling from two pairs of teachers (DeiT-III [62] & DINO [8] and iBOT [81] & dBOT-ft [31][3]), as well as using all four together. In all cases we use publicly available ViT-Base models trained on ImageNet-1K.

Extended protocol. We use the protocol summarized in Sec. 3.2 and detailed in the supplementary material. We additionally report results on ImageNet-v2 [47], an alternative validation set for ImageNet, as well as two datasets for measuring performance under domain shift, *i.e.* ImageNet-R [20] and ImageNet-Sketch [64]. Besides reporting results for all 15 transfer datasets jointly, we further split the datasets into separate axes, *i.e.* for concept generalization [55], long-tail [63] and small-scale fine-grained recognition datasets (Aircraft [35], Cars196 [28], DTD [12], EuroSAT [19], Flowers [38], Pets [40], Food101 [6], SUN397 [69]).

In all cases we chose hyperparameters based on ImageNet-1K performance, the task which corresponds to the distillation data. See the supplementary material for further implementation and evaluation details. There, we further report results using the pre-existing classifiers in a plug-and-play manner, as well as for the case of distillation using synthetic data from the ImageNet-SD dataset [53].

Results. We summarize results for our best UNIC models from different teachers in Figs. 1 and 5. In Fig. 1 we show *relative* gains for a UNIC model trained from all four teachers, while in Fig. 5 we report results for models distilled from three different sets of teachers (DINO & DeiT-III, iBOT & dBOT-ft and all four teachers). A short summary of our most important observations follows.

1. **Stronger teachers give stronger students.** From Fig. 5 we see that iBOT & dBOT-ft yield improved student models compared to DINO & DeiT-III.

[3] We use the dBOT model fine-tuned for ImageNet-1K classification.

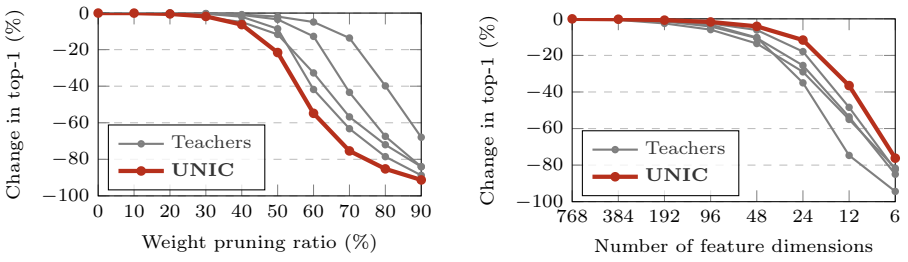
2. **Adding more teachers seems to generally improve performance.** Distilling from all four teachers produces an even stronger student for most cases. This is also true when the additional teachers are not better than the existing ones: Besides ImageNet and transfer, adding DINO & DeiT-III to the ensemble also improves segmentation performance over iBOT.
3. **UNIC models excel at image-level classification**. UNIC from 4 teachers attains 83.8% and 80.3% top-1 accuracy on ImageNet-1K and ImageNet-v2, matching the top performance of the state-of-the-art dBOT-ft model (84% and 80%, respectively). Results are also strong on transfer learning, with UNIC achieving +2.7% higher top-1 on average than iBOT/DINO.
4. **Impressive gains on transfer to small fine-grained datasets**. UNIC achieves a +9.2% *relative gain* on average on 8 small-scale classification datasets, some for domains far outside the ImageNet training set used for all teachers and distillation (*i.e.* including satellite images and textures). Complementary teachers appear to be highly beneficial in this case.
5. **Strong gains for dense prediction with linear probing**. Strong gains are also observed on segmentation and depth estimation, for example on ADE-20K where UNIC achieves a +8.2% relative gain over iBOT. Although far from being the optimal protocol for the task, linear probing is best to evaluate the discriminative power of the patch tokens from the encoder.
6. **Retaining top teacher performance for domain shifts**. DeiT-III shows exceptionally high performance on ImageNet-R and Sketch (51.4% and 39.3% top-1 accuracy, respectively). Our best UNIC model retains this top performance, achieving 51.4% and 38.5%, respectively.

Distilling arbitrary models. We extend our study to larger teachers like MetaCLIP ViT-Huge/14 [70] and DINOv2 ViT-Giant/14 [39] trained on arbitrary datasets. We train a ViT-Large/14 student for 200 epochs at resolution 224, setting the teacher dropping probability to $p = 0.25$. In Tab. 2 we report results for k-NN and zero-shot classification on ImageNet-1K, as well as semantic segmentation on ADE-20K. UNIC* refers to a UNIC model without a ladder of projectors. These results offer some basic verification that our insights are also valid in this more generic distillation case: Our UNIC model outperforms DINOv2 on ImageNet-1K as well as the MetaCLIP on zero-shot classification.

Table 2: Distilling MetaCLIP and DINOv2 on ImageNet-1K.

Model	k-NN	Z-shot	ADE-20K
Teacher Models			
MetaCLIP-H [70]	82.1	80.5	35.4
DINOv2-G [39]	83.4	–	**48.7**
AM-RADIO [46]	84.8	80.4	**48.1**
UNIC*-L	<u>85.0</u>	<u>80.7</u>	<u>47.7</u>
UNIC-L	**85.4**	**81.2**	47.1

Weight and feature space utilization. In this section, we seek to better understand why multi-teacher distillation leads to overall stronger encoders. We do that by investigating the utilization of the encoder weights after pruning (Fig. 6a) and the feature space after dimensionality reduction (Fig. 6b). We report the change in accuracy on ImageNet-1K for our UNIC model and its teachers when we prune the weights or reduce the feature dimension before training lin-



(a) Pruning analysis **(b)** PCA analysis

Fig. 6: Network utility analysis via ImageNet-1K linear probing for the four teachers and our student UNIC distilled from all of them. For each model, before training linear probes, we either **(a)** prune their weights or **(b)** reduce the dimension of their features via PCA. We report change in top-1 accuracy compared to their base performance. UNIC's encoder weights work together more cohesively **(a)**, and its feature space is more robust to dimensionality reduction **(b)**.

ear probes. We prune encoder weights using ℓ_1-norm-based unstructured weight pruning, and perform dimensionality reduction using PCA with whitening.

From Fig. 6a, we see that the performance of UNIC drops more rapidly than any of the teachers as we increase the pruning ratio. This indicates that the encoder weights show improved synergy, working together more cohesively and efficiently to enhance the model's overall performance.

▶ *UNIC encoders utilize weights more effectively*

At the same time, in Fig. 6b, we see that our student preserves its base performance better than all teachers as we reduce the number of dimensions with PCA. It seems that the feature space of UNIC can be represented better with fewer principal components, possibly because of higher entanglement in the original feature space.

▶ *UNIC encoders are more resilient to dimensionality reduction*

5 Conclusions

In this paper, we systematically analyze multi-teacher distillation and introduce improvements to the distillation process that significantly enhance the performance of student models across various benchmarks. More importantly, we show that it is possible to distil from multiple teachers with complementary strengths and learn models that match or improve the respective best teacher in both image- and patch-based classification tasks. In that regard, we view UNIC models as *universal* classification models, advancing the frontier of general representation learning without task-specific adaptation.

Acknowledgements. The authors would like to sincerely thank Myung-Ho Ju, Florent Perronnin, Rafael Sampaio de Rezende, Vassilina Nikoulina and Jean-Marc Andreoli for inspiring discussions and many thoughtful comments.

References

1. Ahn, S., Hu, S.X., Damianou, A., Lawrence, N.D., Dai, Z.: Variational information distillation for knowledge transfer. In: Proc. CVPR (2019)
2. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: Proc. ICKDDM (2019)
3. Asif, U., Tang, J., Harrer, S.: Ensemble knowledge distillation for learning improved and efficient networks. In: Proc. ECAI (2020)
4. Ba, J., Frey, B.: Adaptive dropout for training deep neural networks. In: Proc. NeurIPS (2013)
5. Bhat, S.F., Alhashim, I., Wonka, P.: Adabins: Depth estimation using adaptive bins. In: Proc. CVPR (2021)
6. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101 – Mining discriminative components with random forests. In: Proc. ECCV (2014)
7. Buciluǎ, C., Caruana, R., Niculescu-Mizil, A.: Model compression. In: Proc. SIGKDD (2006)
8. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proc. ICCV (2021)
9. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proc. ICML (2020)
10. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proc. CVPR (2021)
11. Chen, Z., Badrinarayanan, V., Lee, C.Y., Rabinovich, A.: Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In: Proc. ICML (2018)
12. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: Proc. CVPR (2014)
13. Clark, K., Luong, M.T., Khandelwal, U., Manning, C.D., Le, Q.V.: Bam! born-again multi-task networks for natural language understanding. In: ACL (2019)
14. 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: Proc. ICLR (2021)
15. Fukuda, T., Suzuki, M., Kurata, G., Thomas, S., Cui, J., Ramabhadran, B.: Efficient knowledge distillation from an ensemble of teachers. In: Interspeech (2017)
16. Ghiasi, G., Zoph, B., Cubuk, E.D., Le, Q.V., Lin, T.Y.: Multi-task self-training for learning general representations. In: Proc. CVPR (2021)
17. 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. Proc. NeurIPS (2022)
18. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proc. CVPR (2022)
19. Helber, P., Bischke, B., Dengel, A., Borth, D.: EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. JSTAEORS (2019)
20. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., Gilmer, J.: The many faces of robustness: A critical analysis of out-of-distribution generalization. In: Proc. ICCV (2021)
21. Heo, B., Kim, J., Yun, S., Park, H., Kwak, N., Choi, J.Y.: A comprehensive overhaul of feature distillation. In: Proc. ICCV (2019)

22. Heo, B., Lee, M., Yun, S., Choi, J.Y.: Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In: Proc. AAAI (2019)
23. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: Proc. NeurIPS-W (2014)
24. Hu, H., Dey, D., Hebert, M., Bagnell, J.A.: Learning anytime predictions in neural networks via adaptive loss balancing. In: Proc. AAAI (2019)
25. Huang, G., Sun, Y., Liu, Z., Sedra, D., Weinberger, K.Q.: Deep networks with stochastic depth. In: Proc. ECCV (2016)
26. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proc. CVPR (2018)
27. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv:2304.02643 (2023)
28. Krause, J., Deng, J., Stark, M., Li, F.F.: Collecting a large-scale dataset of fine-grained cars. In: Proc. CVPR-W (2013)
29. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Proc. NeurIPS (2012)
30. Landgraf, S., Hillemann, M., Kapler, T., Ulrich, M.: Efficient multi-task uncertainties for joint semantic segmentation and monocular depth estimation. arXiv:2402.10580 (2024)
31. Liu, X., Zhou, J., Kong, T., Lin, X., Ji, R.: Exploring target representations for masked autoencoders. In: Proc. ICLR (2022)
32. Liu, Y., Zhang, W., Wang, J.: Adaptive multi-teacher multi-level knowledge distillation. Neurocomputing (2020)
33. Loshchilov, I., Hutter, F.: SGDR: Stochastic gradient descent with warm restarts. In: Proc. ICLR (2017)
34. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: Proc. ICLR (2019)
35. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv:1306.5151 (2013)
36. Marrie, J., Arbel, M., Mairal, J., Larlus, D.: On good practices for task-specific distillation of large pretrained models. arXiv:2402.11305 (2024)
37. Matena, M.S., Raffel, C.A.: Merging models with fisher-weighted averaging. In: Proc. NeurIPS (2022)
38. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: Proc. ICVGIP (2008)
39. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. TMLR (2024)
40. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.V.: Cats and dogs. In: Proc. CVPR (2012)
41. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in Python. JMLR **12** (2011)
42. Peng, Z., Dong, L., Bao, H., Wei, F., Ye, Q.: A unified view of masked image modeling. TMLR (2023)
43. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Proc. ICML (2021)

44. Ramé, A., Ahuja, K., Zhang, J., Cord, M., Bottou, L., Lopez-Paz, D.: Model ratatouille: Recycling diverse models for out-of-distribution generalization. In: Proc. ICML (2023)
45. Rame, A., Kirchmeyer, M., Rahier, T., Rakotomamonjy, A., Gallinari, P., Cord, M.: Diverse weight averaging for out-of-distribution generalization. In: Proc. NeurIPS (2022)
46. Ranzinger, M., Heinrich, G., Kautz, J., Molchanov, P.: AM-RADIO: Agglomerative model–reduce all domains into one. In: Proc. CVPR (2024)
47. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do ImageNet classifiers generalize to ImageNet? In: Proc. ICML (2019)
48. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proc. CVPR (2022)
49. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. In: Proc. ICLR (2015)
50. Roth, K., Milbich, T., Ommer, B., Cohen, J.P., Ghassemi, M.: Simultaneous similarity-based self-distillation for deep metric learning. In: Proc. ICML (2021)
51. Roth, K., Thede, L., Koepke, A.S., Vinyals, O., Henaff, O.J., Akata, Z.: Fantastic gains and where to find them: On the existence and prospect of general knowledge transfer between any pretrained model. In: Proc. ICLR (2024)
52. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. IJCV **115**(3) (2015)
53. Sariyildiz, M.B., Alahari, K., Larlus, D., Kalantidis, Y.: Fake it till you make it: Learning transferable representations from synthetic ImageNet clones. In: Proc. CVPR (2023)
54. Sariyildiz, M.B., Kalantidis, Y., Alahari, K., Larlus, D.: No reason for no supervision: Improved generalization in supervised models. In: Proc. ICLR (2023)
55. Sariyildiz, M.B., Kalantidis, Y., Larlus, D., Alahari, K.: Concept generalization in visual representation learning. In: Proc. ICCV (2021)
56. Shi, B., Zhang, X., Wang, Y., Li, J., Dai, W., Zou, J., Xiong, H., Tian, Q.: Hybrid distillation: Connecting masked autoencoders with contrastive learners. In: Proc. ICLR (2024)
57. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: Proc. ECCV (2012)
58. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. JMLR **15**(1) (2014)
59. Stoica, G., Bolya, D., Bjorner, J., Ramesh, P., Hearn, T., Hoffman, J.: Zipit! merging models from different tasks without training. In: Proc. ICLR (2024)
60. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. In: Proc. ICLR (2020)
61. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: Proc. ICML (2021)
62. Touvron, H., Cord, M., Jegou, H.: DeiT III: Revenge of the ViT. In: Proc. ECCV (2022)
63. 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: Proc. CVPR (2018)
64. Wang, H., Ge, S., Lipton, Z., Xing, E.P.: Learning robust global representations by penalizing local predictive power. In: Proc. NeurIPS (2019)

65. Wang, H., Vasu, P.K.A., Faghri, F., Vemulapalli, R., Farajtabar, M., Mehta, S., Rastegari, M., Tuzel, O., Pouransari, H.: SAM-CLIP: Merging vision foundation models towards semantic and spatial understanding. In: Proc. CVPR-W (2023)
66. Wang, Y., Tang, S., Zhu, F., Bai, L., Zhao, R., Qi, D., Ouyang, W.: Revisiting the transferability of supervised pretraining: an MLP perspective. In: Proc. CVPR (2022)
67. Wei, L., Xie, L., Zhou, W., Li, H., Tian, Q.: Mvp: Multimodality-guided visual pre-training. In: Proc. ECCV (2022)
68. Wortsman, M., Ilharco, G., Gadre, S.Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A.S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., et al.: Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In: Proc. ICML (2022)
69. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: SUN database: Large-scale scene recognition from abbey to zoo. In: Proc. CVPR (2010)
70. Xu, H., Xie, S., Tan, X.E., Huang, P.Y., Howes, R., Sharma, V., Li, S.W., Ghosh, G., Zettlemoyer, L., Feichtenhofer, C.: Demystifying clip data. In: Proc. ICLR (2024)
71. Yao, Y., Desai, N., Palaniswami, M.: MOMA: Distill from self-supervised teachers. arXiv:2302.02089 (2023)
72. Ye, P., Huang, C., Shen, M., Chen, T., Huang, Y., Zhang, Y., Ouyang, W.: Merging vision transformers from different tasks and domains. arXiv:2312.16240 (2023)
73. Yim, J., Joo, D., Bae, J., Kim, J.: A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In: Proc. CVPR (2017)
74. Yin, D., Han, X., Li, B., Feng, H., Bai, J.: Parameter-efficient is not sufficient: Exploring parameter, memory, and time efficient adapter tuning for dense predictions. arXiv:2306.09729 (2023)
75. You, S., Xu, C., Xu, C., Tao, D.: Learning from multiple teacher networks. In: Proc. SIGKDD (2017)
76. Ypsilantis, N.A., Chen, K., Araujo, A., Chum, O.: Udon: Universal dynamic online distillation for generic image representations. arXiv 2406.08332 (2024)
77. Yuan, H., Li, X., Zhou, C., Li, Y., Chen, K., Loy, C.C.: Open-vocabulary SAM: Segment and recognize twenty-thousand classes interactively. In: Proc. ECCV (2024)
78. Yun, H., Cho, H.: Achievement-based training progress balancing for multi-task learning. In: Proc. ICCV (2023)
79. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: Proc. ICLR (2017)
80. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ADE20k dataset. IJCV (2019)
81. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: iBOT: Image BERT pre-training with online tokenizer. In: Proc. ICLR (2022)