

Aggregating Cross-Domain Knowledge via Learnable Tokens for Multi-Teacher Distillation

Wu Ran¹, Weijia Zhang¹, Shuyang Pang², Jingsheng Liu², Xiaohui Zhang²,
Yichao Yan¹, and Chao Ma^{*,1}

¹ MoE Key Lab of Artificial Intelligence, Institute of AI, Shanghai Jiao Tong
University, China

{bonjourlemonde, weijia.zhang, yanyichao, chaoma}@sjtu.edu.cn

² CISDI Information Technology CO., LTD., China

{shuyang.pang, jingsheng.liu, xiaohui.zhang}@cisdi.com.cn

Abstract. Multi-teacher distillation has emerged as a powerful paradigm for consolidating diverse visual knowledge from multiple Vision Foundation Models (VFMs). However, existing methods typically rely on simultaneous feature alignment, which forces the student to match multiple teacher distributions directly. This rigid approach often induces representational dissonance and semantic conflicts, where teachers with disparate architectures and divergent optimization goals lead to sub-optimal performance and restricted architectural flexibility. In this paper, we propose a novel framework termed ACTok that recasts multi-teacher distillation as a token-mediated aggregation process. By treating visual features from both CNNs and Transformers as sets of tokens, we introduce learnable tokens that serve as flexible neural proxies to bridge diverse representational spaces. To avoid specialized modifications on student backbones, we further design an alternating interaction mechanism that enables these learnable tokens to dynamically reconcile complementary knowledge across teachers. Our ACTok is inherently architecture-agnostic, facilitating effective knowledge transfer to both CNN and ViT students. Furthermore, by resolving teacher conflicts at the token level, our method achieves superior performance using only the ImageNet-1K dataset, rendering it orders of magnitude more data-efficient than existing billion-scale distillation baselines. Extensive experiments on classification, segmentation, and depth estimation demonstrate that our token-based aggregation consistently outperforms state-of-the-art multi-teacher distillation methods, providing a robust and scalable solution for VFM knowledge consolidation. Code is publicly available on <https://github.com/VISION-SJTU/ACTok>.

Keywords: Vision foundation models · Multi-domain knowledge aggregation · Learnable tokens

* Corresponding author.

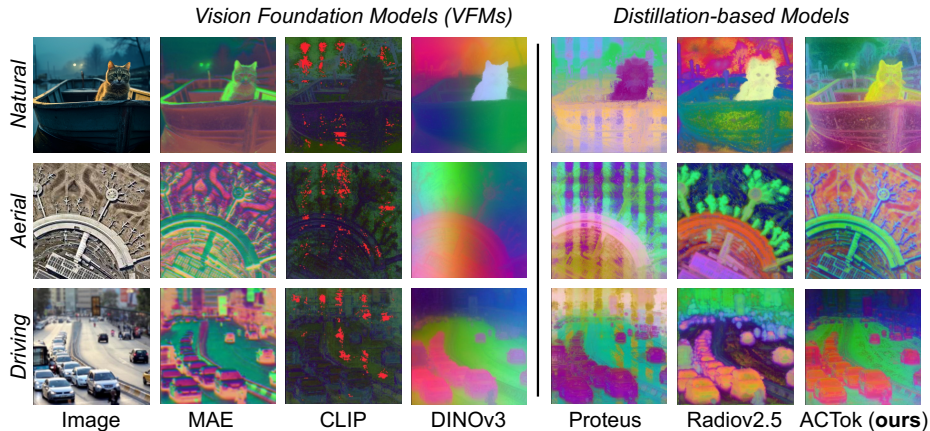


Fig. 1: PCA visualization shows that ACTok provides rich fine-grained semantic details for perception. We compare ACTok with VFMs: MAE [21], CLIP [42], DINOv3 [50], and current distillation-based approaches, Proteus [64], and Radiov2.5 [24]. We choose the ViT-B variant for all methods for PCA visualization. We employ PCA to project the features into RGB space for visualization.

1 Introduction

Vision Foundation Models (VFMs) [7, 14, 21, 40, 45] have fundamentally shifted the computer vision landscape by providing robust representations that generalize across a vast spectrum of downstream tasks. To consolidate the complementary strengths of diverse VFMs [21, 31, 50], recent research has turned towards multi-teacher knowledge distillation [6, 15, 24, 44, 62, 68]. The ultimate goal is to develop a unified, versatile student that inherits a collective intelligence aggregated from multiple VFM experts.

Despite their potential, current multi-teacher distillation paradigms face a fundamental challenge: the representational dissonance among heterogeneous teachers. For instance, MAE [21] and DINOv3 [50] present distinct discrepancies in global and detail perception, as illustrated in Figure 1. Most existing frameworks [6, 24, 44, 64] enforce the student to perform direct, simultaneous feature alignment with multiple teachers. As illustrated in Figure 2(b), this one-to-many matching strategy assumes that different teachers share a compatible representational space. Moreover, as shown in Figure 2(a), different VFMs [21, 50] often focus on distinct and sometimes contradictory image properties due to their divergent pre-training objectives. Naïvely forcing a student’s backbone to align with multiple heterogeneous targets simultaneously induces semantic conflicts and gradient interference, where signals from one teacher may wash out the nuances of another. To counteract these conflicts, existing methods often resort to brute-force consensus through billion-scale proxy datasets (*e.g.*, DataComp-1B [18]), rendering the distillation process computationally prohibitive and restricted to specific architectures like Vision Transformers [14].

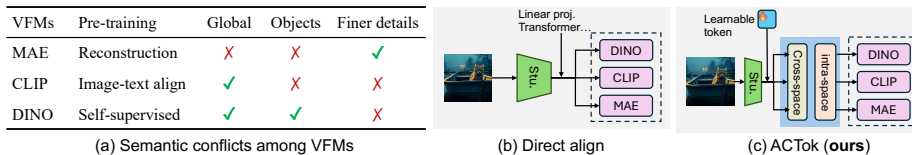


Fig. 2: Current VFMs [21, 42, 50] exhibit conflicting semantic preferences across global, object-level, and fine-grained perception. Directly aligning students with VFMs like (b) may cause teacher competition. (c) introduces ACTok to aggregate cross-domain knowledge at the token level.

In this paper, we propose a novel framework, dubbed ACTok, which recasts multi-teacher distillation as a token-mediated aggregation process. Our approach is grounded in the unified perspective that visual features from either Convolutional Neural Networks (CNNs) [29] or Transformers [14] can be regarded as sets of tokens. Inspired by the success of using tokens to aggregate multi-view information [55] or unify grounding representations [46], we introduce learnable tokens that act as flexible neural proxies to aggregate cross-domain visual knowledge embedded in diverse teachers. To avoid specialized and intrusive modifications to the student’s backbone, we develop an alternating interaction mechanism positioned behind the backbone. This mechanism allows the learnable tokens to dynamically aggregate and reconcile complementary knowledge across multiple representational spaces via alternating cross-space and intra-space modules as illustrated in Figure 2(c), which facilitates absorbing cross-space knowledge while consolidating internal visual features.

As a consequence, ACTok enables effective multi-teacher knowledge distillation using only the ImageNet-1k [12] dataset, supporting flexible student-teacher pairs regardless of their specific architectures. By resolving teacher conflicts at the token level, ACTok is inherently architecture-agnostic and data-efficient. Extensive experiments across four diverse vision tasks: ImageNet-1k classification, semantic segmentation, fine-grained recognition, and depth estimation, demonstrate that ACTok significantly enhances the student’s generalization. As shown in Figure 1(a), our distilled student captures richer semantic cues and finer details than individual teachers. Our main contributions are summarized as follows:

- We identify the representational dissonance in existing multi-teacher distillation and propose to recast the task as a token-mediated aggregation process to mitigate semantic conflicts.
- We introduce learnable tokens and an alternating interaction mechanism (cross-space and intra-space) that dynamically reconciles heterogeneous teacher knowledge while maintaining architectural flexibility for both CNN and ViT students.
- We demonstrate that our framework is remarkably data-efficient, outperforming state-of-the-art methods on multiple downstream benchmarks while using significantly less proxy data, providing a robust solution for consolidating VFM knowledge.

2 Related Work

Vision Foundation Models (VFMs). The landscape of computer vision has been revolutionized by VFMs trained on massive datasets. Early self-supervised methods like MoCo [22] and SimCLR [8] focus on instance discrimination, while DINO [7] introduced self-distillation to capture local-to-global structures. Masked Image Modeling (MIM) approaches, such as MAE [21] and BEiT [4] draw inspiration from [13] and emphasize spatial reconstruction, leading to strong pixel-level awareness. Concurrently, multi-modal models like CLIP [42] and ALIGN [28] align visual and textual spaces for zero-shot reasoning, while SAM [31] provides a universal backbone for promptable segmentation. Recently, DINOv3 [50] further scales these capabilities. Crucially, these models exhibit significant representation gaps due to their diverse objectives (*e.g.*, contrastive vs. classification) and architectures (*e.g.*, CNN vs. ViT). Our work focuses on bridging these gaps by aggregating their complementary strengths into a single, unified student.

Knowledge Distillation. Knowledge distillation (KD) [3, 5, 9, 19, 26, 30, 39] typically transfers knowledge from a large teacher to a smaller student by matching logits or intermediate features [1, 25, 27, 47, 56, 61]. While recent works like MobileSAM [62], TinyCLIP [57], and Proteus [64] achieve impressive results, they are largely constrained to one-to-one teacher-student pairs within the same architectural family. On the other hand, cross-architecture distillation [20, 35, 63] remains challenging as it often requires hand-crafted projection layers to handle dimension or resolution mismatches. Unlike traditional KD that treats features as targets for direct alignment, we introduce learnable tokens to dynamically digest teacher-specific knowledge across heterogeneous architectures, ensuring a more flexible and robust transfer process.

Multi-teacher Distillation for VFMs. Hinton *et al.* [26] explore the concept of ensemble distillation. Later, ensemble distillation studies [2, 17, 34, 58–60, 65, 69, 70] mainly concentrate on knowledge aggregation from conventional vision networks. Recently, Ranzinger *et al.* [44] introduce an agglomerative distillation approach to reducing the domain knowledge embedded in various vision foundation models into a single student model. Through a direct global-level and patch-level alignment with CLIP [42], SAM [31], and DINOv2 [41], the ultimate student model is shown to inherit rich knowledge for various vision-related tasks. Soon, RADIO-v2.5 [24] extends this concept with multi-resolution distillation, which facilitates high-resolution recognition. DUNE [48] further extends the alignment projection with Transformers, facilitating knowledge aggregation among 2D and 3D domains. However, they directly impose feature alignment between the student and multiple teachers, risking unsatisfactory optimization caused by semantic conflicts among VFMs. In contrast, our ACTok framework shifts the paradigm toward token-mediated aggregation since visual features from CNNs and ViTs can both be regarded as tokens. With this purpose, we resolve teacher conflicts at the token level through introducing learnable tokens, presenting an inherent architecture-agnostic approach.

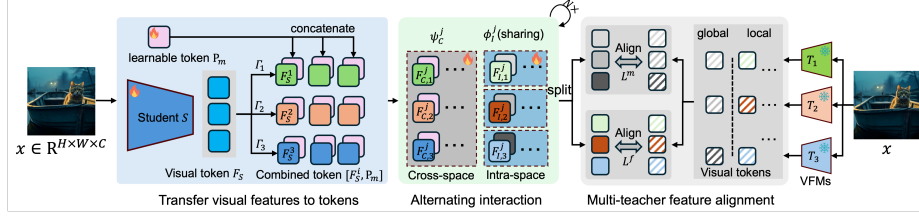


Fig.3: Schematic illustration of the proposed ACTok framework. Our approach aggregates multi-domain knowledge from diverse VFM teachers into a unified student through a token-mediated aggregation process. A learnable token P_m acts as a neural proxy, interacting with the student’s projected features across M teacher-specific representational spaces. Finally, the student’s backbone is optimized via dual-level alignment: global-level alignment via the learnable token and patch-level alignment.

3 Methodology

In this section, we formally introduce ACTok, a token-mediated framework for multi-teacher knowledge distillation. We first provide an overview of the reformalized distillation paradigm, followed by a detailed description of the alternating interaction mechanism and alignment process.

3.1 Overview

Definition. We consider a set of M pre-trained vision foundation model (VFM) teachers $\mathcal{T} = \{T_i\}_{i=1}^M$, which have heterogeneous architectures (*e.g.*, CNN, ViT) and diverse pre-training objectives. Our goal is to train a single student model S that consolidates the collective intelligence of these teachers. Unlike prior works that rely on billion-scale datasets, we instead employ ImageNet-1k as a proxy dataset $\mathcal{D}$ to achieve high data efficiency.

Transfer Visual Features into Tokens. For a given image $x \in \mathbb{R}^{H \times W \times C}$, the student model S and teachers T_i extract visual features F_S and F_{T_i} , respectively. To bridge the architectural gap between CNNs and Transformers, we recast all visual features as sets of tokens. Specifically, for a CNN-based model, the feature map $F \in \mathbb{R}^{H \times W \times d}$ is viewed as a collection of $H \times W$ spatial tokens, each of dimension d . For a ViT-based model, the feature $F \in \mathbb{R}^{(1+L) \times d}$ naturally consists of a class token and L patch tokens. This unified token-based representation allows our framework to be inherently architecture-agnostic. For the purpose of aligning F_S with teacher features F_{T_i} , we need to project it onto M teacher-specific spaces, *i.e.*, $F_S^i = \Gamma_i(F_S)$, where Γ_i denotes MLP projection [24] or Transformer block [48]. Intuitively, forcing the alignment between each F_S^i with F_{T_i} results in knowledge transfer from VFMs to the student. To avoid dimension mismatch, we set the dimension of F_S^i to $d_T^{\max} = \max\{d_{T_1}, d_{T_2}, \dots, d_{T_M}\}$.

Alternating Interaction. As shown in the left of Section 3.2, simply aligning F_S^i with F_{T_i} results in a degraded student suffering from the semantic conflicts

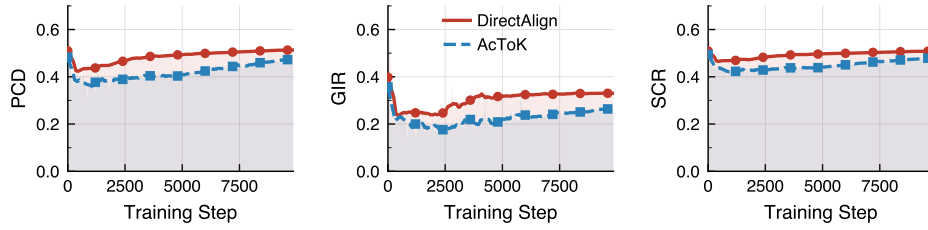


Fig. 4: Directly aligning multiple teachers causes severe gradient conflict throughout the distillation period. The results are obtained using a ViT-B student distilled from CLIP, MAE, and DINOv3 ViT-B teachers on 1% ImageNet.

among VFMs. An in-depth analysis is provided in Figure 4, where we employ three metrics, i.e., PCD, GIR, and SCR (please refer to supplementary material for details) to measure the gradient conflict among teachers during training process. The Direct Align approach tends to suffer from heavy gradient conflict among teachers, thereby confusing the student on what to learn. To mitigate this representational dissonance among heterogeneous teachers, we introduce a learnable token $P_m \in \mathbb{R}^{d_m}$ as a neural proxy. Instead of direct backbone-to-backbone alignment, AcTok decouples the distillation process via an alternating interaction mechanism. In each of the N stages, the learnable token first performs *cross-space interaction* to aggregate complementary knowledge from all M projected feature spaces. Subsequently, *intra-space interaction* is conducted to consolidate the assimilated information within each specific teacher’s domain. This mechanism ensures that the conflicts between divergent teacher targets are alleviated at the token level before the knowledge is integrated into the student. Details will be provided in Section 3.2.

Alignment with VFMs. To facilitate aggregating cross-domain knowledge from different VFMs, we further develop a dual-level alignment strategy where the learnable token is responsible for assimilating global-level knowledge from all VFMs with the least semantic conflicts. We elaborate this strategy in Section 3.3.

3.2 Alternating Interaction Mechanism

By regarding visual features from CNNs and ViTs as a set of tokens, we transfer the multi-teacher distillation problem into a token-mediated cross-domain knowledge aggregation process. This is inspired by the success of recent work [46, 55] that aggregates multi-view information and grounding representations with learnable tokens, respectively.

Starting from the visual features from the student’s backbone, noted as F_S^i , we introduce a learnable token $P_m \in \mathbb{R}^{d_m}$ that dynamically interacts with all F_S^i in M representational spaces. Specifically, we first design a cross-space module ψ to explore cross-domain consensus knowledge by operating on the combination of all F_S^i and P_m . Then, we introduce an intra-space module ϕ that operates *separately* on each representative space to consolidate space-specific knowledge,

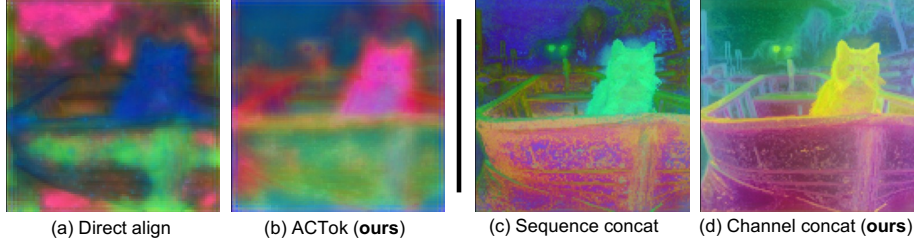


Fig. 5: PCA visualization comparison between ACTok and existing approaches. *Left:* Directly aligning the student with multiple VFMs corrupts the semantic structure. ConvNeXt-B student distilled from three ConvNeXt-B teachers is selected for comparison. *Right:* Comparison of the effectiveness of feature concatenation using ViT-B variants.

the output of which could be aligned with a specific teacher. Through N alternating cross-space and intra-space interaction processes, the ultimate feature in each representational space could be safely utilized for the alignment with the corresponding VFM teacher. As a consequence, the learnable token P_m could consecutively integrate cross-domain knowledge within the alternating interaction procedure. Denote $F_{C,i}^j$ and $F_{I,i}^j$ ($i = 1, \dots, M$) as the output of j -th cross-space and intra-space modules, corresponding to the i -th representational space, respectively. This alternating interaction process can be formulated as

$$\begin{aligned} F_C^j &= \psi_j([F_{I,1}^{j-1}, \dots, F_{I,M}^{j-1}]), \\ F_{C,i}^j &= \text{split}(F_C^j), \\ F_{I,i}^j &= \phi_{j,i}(F_{C,i}^j), \quad j = 1, \dots, N; \quad i = 1, \dots, M, \end{aligned} \quad (1)$$

where $\text{split}(\cdot)$ indicates a channel splitting operation, $[\dots]$ represents channel concatenation. ψ_j denotes the j -th cross-space module, while $\phi_{j,i}$ is the j -th intra-space module corresponding to the i -th representational space. Notably, we set $F_{I,i}^0 = [F_S^i, P_m]$, *i.e.*, the concatenation of the projected student visual feature and the learnable token, before the alternating interaction.

Cross-space Interaction Module. It is noteworthy that the concatenation results in a total number of $M \times (d_T^{\max} + d_m)$ channels, scaling linearly with the number of teachers. Processing this concatenated features with vanilla convolution for CNN or attention for ViT would incur unacceptable computation cost. Hence, we adopt the following techniques to reduce the computation burden.

For CNN students, inspired by [36], we construct the module with a Conv-GELU-Conv design with residual connection [23], where the first convolution is a depth-wise convolution while the latter is a point-wise convolution with groups linearly scaling with M . For the ViT student, the cross-space module is constructed using a vanilla attention block in [14], where its head number is enlarged linearly to the teacher quantity.

Intra-space Interaction Module. As mentioned above, the intra-space module operates independently on each projected feature, each containing $d_T^{\max} + d_m$

channels. Therefore, we share the parameters of $\phi_{j,i}$ ($i = 1, \dots, M$), *i.e.*, $\phi_{j,i} = \phi_j$. This not only reduces the trainable parameters but also enables data-parallel computation for all M features $F_{C,i}^j$ (Equation (1)). This is realized by absorbing M into the batch dimension before passing to ϕ_j . Analogously, the intra-space is implemented with a residual Conv-GELU-Conv block for the CNN student, while a vanilla attention block for the ViT student.

Discussion. Different from representative work [55] which concatenates the learnable token at the sequence dimension, our ACTok employs channel concatenation. First, visual tokens from multi-view share potential 3D correspondence, while tokens from VFMs potentially have semantic conflicts. Concatenating these heterogeneous tokens along the sequence dimension may pollute the attention weights, degrading distillation performance. As shown in the right of Section 3.2, distilling a ViT-B using sequence concatenation produces more noise and artifacts. Second, sequence concatenation is prohibitive for CNN architecture, which requires a rectangular spatial resolution. Last, sequence concatenation incurs $\mathcal{O}(M^2L^2)$ complexity, while we can flexibly reduce the complexity incurred by channel concatenation through group convolution and enlarging attention heads.

3.3 Multi-teacher Feature Alignment

Current multi-teacher distillation approaches like [64], and [24] harshly align the student features with different teachers at multiple levels, risking representation collision among teachers. We argue that a careful alignment between the student and the teachers must be taken into consideration.

Global-level Alignment via Learnable Tokens. We leverage the *learnable tokens* to perform global-level alignment across all teachers. Given the extensive interaction within multiple representational spaces, this alignment acts as a high-level semantic constraint that forces the tokens to aggregate the consensus knowledge across heterogeneous teachers. Specifically, by aligning the same set of learnable tokens with multiple VFM experts, we encourage the tokens to reconcile teacher-specific discrepancies and distill a unified, model-agnostic representational space.

As illustrated in Figure 3, we first derive the ultimate features $F_{L,i}^N$ and decouple them via channel-wise splitting into the student feature $P_i^{\text{fea}} \in \mathbb{R}^{\dots \times d_T^{\text{max}}}$ and the aggregated token $P_i^{\text{tok}} \in \mathbb{R}^{\dots \times d_m}$. To obtain a unified global representation P_i^{glo} , we extract the **cls** token for ViT-based architectures or apply global average pooling for CNN-based backbones on P_i^{tok} , respectively. The learnable token alignment loss is thus formulated as:

$$\mathcal{L}_i^m = \left\| \varphi(P_i^{\text{glo}}) - F_{T_i}^{\text{glo}} \right\|_F^2, \quad (2)$$

where $\varphi(\cdot)$ is a shared linear transformation. $F_{T_i}^{\text{glo}}$ represents a global token for teacher T_i , which is extracted from the **cls** token of F_{T_i} (ViT teacher) or its average pooling result (CNN teacher).

Patch-level Alignment. In addition to global consensus, we align the decoupled student features P_i^{fea} with their corresponding teachers $\{T_i\}$. Given that VFMs are optimized with divergent objectives, their output features reside in distinct manifolds. To honor these discrepancies, we employ a teacher-specific alignment strategy. Formally, the feature alignment loss $\mathcal{L}_i^f$ is defined as:

$$\mathcal{L}_i^f = \begin{cases} 1 - \cos(\text{Pool}(\delta_i(P_i^{\text{fea}})), F_{T_i}), & \text{if } T_i = \text{CLIP}, \\ 1 - \cos(\delta_i(P_i^{\text{fea}}), F_{T_i}), & \text{if } T_i = \text{DINO}, \\ \|\delta_i(P_i^{\text{fea}}) - F_{T_i}\|_F^2, & \text{otherwise,} \end{cases} \quad (3)$$

where $\delta_i(\cdot)$ is a linear projector to match dimension, and $\text{Pool}(\cdot)$ is spatial average pooling. $\cos(\cdot, \cdot)$ computes the cosine similarity between two features.

Training Objective. The overall objective of ACTok is to minimize the joint loss, facilitating both global consensus and local knowledge inheritance:

$$\mathcal{L} = \sum_{i=1}^M \mathcal{L}_i^m + \sum_{i=1}^M \mathcal{L}_i^f, \quad (4)$$

In our experiments, we assign equal weights to all teachers to maintain a balanced aggregation of cross-domain knowledge.

Spatial Resolution Alignment. A technical challenge arises from the *spatial mismatch* between hybrid CNN-ViT pairs or teachers with inconsistent patch counts. To ensure valid computation for $\mathcal{L}_i^f$, we first reshape both P_i^{fea} and F_{T_i} into 4D tensors of shape $B \times C \times H \times W$ (where $H = W = \sqrt{L}$ for ViTs). Subsequently, we employ bilinear interpolation to downsample the higher-resolution feature map to match the spatial dimensions of the smaller one.

3.4 Downstream Adaptation

It is noteworthy that the multi-space interaction modules in ACTok embed high-density cross-domain knowledge consolidated from multiple VFM teachers. Such sophisticated representational structures can be susceptible to catastrophic forgetting or semantic distortion when fine-tuned on limited downstream data. To mitigate this, we apply a smaller learning rate scaling factor (*e.g.*, $0.1\times$) for them during downstream adaptation.

4 Experiments

4.1 Experimental Setup

Implementation Details. We conduct multi-teacher distillation on various student architectures, including ResNet-50 [23], ConvNeXt-B [36], ViT-B/16 [14], and ViT-L/16 [14]. For each *randomly* initialized student, we employ its pre-trained classification-based (CLIP for ViT), MAE-based, and DINO-based counterparts as teachers. For DINO-based teachers, we adopt the latest DINOv3

Table 1: A comprehensive comparison across ImageNet classification, fine-grained classification, image segmentation, and depth estimation with ViT student. † denotes methods for which official results are reported. Best distillation performance is noted in **bold font**.

Method	Student Arch.	Teacher Arch.	# imgs	ImageNet	Fine-grained Cls.			Segmentation ADE20K probing (mIoU†)	Depth Estimation NYUv2 (RMSE↓)
					Aircraft	Caltech101	CUB		
CLIP [42]	ViT-B/16	-	400M	84.2	82.1	90.7	85.1	-	-
MAE [21]	ViT-B/16	-	1.2M	83.6	85.9	91.6	81.2	22.2	0.347
DINOv3 [50]	ViT-B/16	-	1689M	86.2	92.6	93.5	90.3	50.9	0.293
Proteus† [64]	ViT-B/14	ViT-G	1.2M	84.9	91.0	91.9	90.5	-	0.304
RADIOv2.5† [24]	RADIOv2.5-B	ViT-H	~1400M	84.8	83.5	91.2	85.9	48.9	-
ACTok (ours)	ViT-B/16	ViT-B	1.2M	85.6	91.1	94.4	90.8	46.5	0.300
CLIP [42]	ViT-L/14	-	400M	86.7	84.4	91.4	88.2	-	-
MAE [21]	ViT-L/16	-	1.2M	85.9	91.5	92.4	85.5	27.1	0.267
DINOv3 [50]	ViT-L/16	-	1689M	88.4	94.5	93.7	91.5	54.6	0.224
Proteus† [64]	ViT-L/14	ViT-G	1.2M	86.2	93.5	92.1	91.1	-	0.240
RADIOv2.5† [24]	RADIOv2.5-L	ViT-H	~1400M	87.4	86.3	92.5	88.2	51.5	-
ACTok (ours)	ViT-L/16	ViT-L	1.2M	87.5	92.1	93.8	91.4	50.3	0.258

ConvNeXt-B and ViT series. Distillation training is conducted on the ImageNet-1K dataset following a standard 300-epoch training configuration with 224×224 image resolution [36, 53]. We use two cross-space and intra-space modules for the CNN student, while one for the ViT student, separately. We set the dimension of the learnable token to 1024 for ResNet-50, 512 for ConvNeXt-B students. For ViT-B/16 and ViT-L/16 students, we set the learnable token’s dimension to 512 and 768, respectively. The batch size for CNNs and ViT-B/16 students is fixed as 1024. The batch size is reduced to 512 for distilling ViT-L/16 student due to GPU memory constraints. All experiments are conducted on four NVIDIA A800 80G GPUs, utilizing AdamW [37] optimizer with learning rate $1e-4$ and weight decay 0.01. We adopt mixed precision training. More details on the training configurations and the architecture are provided in the supplementary material.

Benchmarks. We evaluate the model’s capability across various vision tasks and datasets: image classification on ImageNet-1K [12], fine-grained image classification on Aircraft [38], Caltech101 [16], and CUB2011 [54] datasets, semantic segmentation on ADE20K [67], and depth estimation on NYUv2 [49]. We first fine-tune all listed models in Tables 1 and 2 except for Proteus and RADIO-v2.5 on the ImageNet-1k dataset for 10 epochs and report the evaluation top-1 accuracy. We then fine-tune these models on three fine-grained datasets with a batch size of 32 for 100 epochs. The input size is set to 224×224 for both ImageNet and fine-grained classification tasks. For the ADE20K benchmark, we conduct linear probing while freezing the backbone following [50] to compare the feature quality. We train models with a resolution of 512×512 and report the mean Intersection over Union (mIoU) metric. Following [64], we adopt the DPT [43] decoder for depth estimation on NYUv2. Typically, we employ a batch size of 16 for 25k iterations under the MMSegmentation [10] framework. The Root Mean Square Error (RMSE) metric is employed for quantitative comparison. More details can be found in the supplementary material.

Table 2: A comprehensive comparison across ImageNet classification, fine-grained classification, image segmentation, and depth estimation with CNN student. † denotes methods for which official results are reported.

Method	Student Arch.	Teacher Arch.	# imgs	ImageNet	Fine-grained Cls.			Segmentation	Depth Estimation
					Aircraft	Caltech101	CUB		
Classification [23]	-	-	1.2M	80.9	81.6	90.5	83.3	12.4	0.439
MAE [21]	ResNet-50	-	1.2M	65.0	83.6	88.0	71.7	15.8	0.428
DINO [7]		-	1.2M	80.0	76.1	89.6	74.8	18.4	0.426
ACTok (ours)		ResNet-50	1.2M	78.0	85.4	89.9	79.7	20.5	0.408
Classification [36]	-	-	1.2M	83.8	87.0	92.9	90.6	18.0	0.355
Spark [52]	-	-	1.2M	84.8	92.1	91.6	82.6	31.3	0.329
DINOv3 [50]	ConvNeXt-B	-	1689M	86.3	94.0	92.7	89.1	36.4	0.293
ACTok (ours)		ConvNeXt-B	1.2M	83.0	88.7	93.1	88.7	34.6	0.361
ACTok (ours)		ViT-B	1.2M	84.1	90.0	93.5	88.7	34.7	0.349

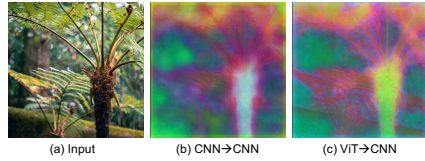


Fig. 6: PCA visualization comparison on hybrid distillation with ConvNeXt-B student.

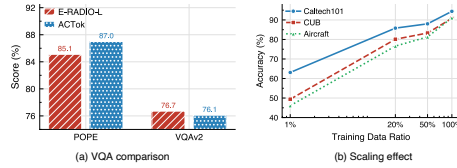


Fig. 7: Scaling effect and comparison on multimodal VQA task.

4.2 Main Results

We examine the capability of the ACTok-distilled student across various downstream vision tasks by comparing their performance with the corresponding teachers. Additionally, we select two recent advanced distillation approaches, Proteus [64] and RADIO-v2.5 [24] for comparison. In particular, we collect the officially reported metrics for these two methods for a fair comparison.

Multi-teacher Distillation into Transformers. We distill the official CLIP ViT-B/16, MAE-pretrained ViT-B/16, and DINOv3 ViT-B/16 models into a vanilla ViT-B/16 student. We do not adopt the advanced 2D rope position embedding [51], hybrid architecture design like [24], and register tokens [11] that could benefit the student’s performance. Table 1 presents quantitative results.

We observe that the student achieves impressive performance on ImageNet and fine-grained classification. Notably, on Caltech101, the student even surpasses all teachers. On dense prediction tasks, the ViT-B/16 student consistently outperforms the MAE teacher with a 24.3% linear probing mIoU gain on semantic segmentation and a 0.047 RMSE improvement on depth estimation.

The distilled student also outperforms Proteus-B on ImageNet classification, fine-grained classification, and depth estimation. On the other hand, since the distilled students tend to provide features with distinct details, linear probing may not unleash its potential in dense segmentation, which requires continuous feature maps. Nonetheless, with vision knowledge aggregated from other teachers, the student even outperforms DINOv3 on the Caltech101 and CUB datasets. Note that our method only falls slightly short of RADIO-v2.5 on semantic seg-

Table 3: Multi-teacher distillation using teachers of hybrid architectures. We use different combinations of teacher models to distill a ConvNeXt-B student. The best results are in **bold**.

Combination	Teacher-I (CLS)	Teacher-II (MAE)	Teacher-III (DINO)	$\mathcal{L}_1^m + \mathcal{L}_2^m + \mathcal{L}_3^m$	$\mathcal{L}_1^f + \mathcal{L}_2^f + \mathcal{L}_3^f$
Homogeneous	ConvNeXt-B	ConvNeXt-B	ConvNeXt-B	0.20	0.23
Hybrid	CLIP	ConvNeXt-B	ConvNeXt-B	0.11	0.27
Hybrid	ConvNeXt-B	ViT-B/16	ConvNeXt-B	0.18	0.48
Hybrid	CLIP	ViT-B/16	ViT-B/16	0.09	0.08

mentation, while significantly surpassing it on all fine-grained datasets (*e.g.* 7.6% accuracy gain on Aircraft with ViT-B student).

Multi-teacher Distillation into CNNs. We further adopt ACTok to distill two CNN variants, *i.e.*, ResNet-50 and ConvNeXt-B, where their ImageNet classification pre-trained, MAE, and DINO counterparts are used as teachers. Quantitative results are presented in Table 2. We observe that the ResNet-50 student achieves substantial improvements on dense prediction tasks, including segmentation and depth estimation, whereas the ConvNeXt-B student struggles to integrate knowledge from VFMs. We conjecture that the strong inductive bias of CNN with dataset bias will amplify the difficulty for multi-teacher distillation among CNNs. We show that hybrid multi-teacher distillation alleviate this.

Multi-teacher Distillation from Hybrid Teacher Architectures. Cross-architecture distillation entails significant challenges in reconciling heterogeneous spatial priors [20, 63]. Nevertheless, as evidenced in Table 1, distilling a ConvNeXt-B student using an *all-ViT* teacher ensemble consistently yields superior performance across diverse tasks. This leads to a counter-intuitive observation: *In the presence of substantial dataset bias between the proxy and pre-training datasets, homogeneous ViT teachers effectively alleviate performance degradation for CNN-based students.*

We investigate this phenomenon by fixing the student as ConvNeXt-B while varying the teacher architectures in Table 3. While substituting a CNN teacher with CLIP-ViT marginally reduces training loss, a heterogeneous mix (*e.g.*, ViT-based MAE and CNN-based DINO) complicates the optimization due to conflicting spatial feature alignments. Conversely, an all-ViT ensemble facilitates a smoother distillation process. We posit that ViTs, characterized by weaker inductive biases and finer spatial resolutions ($16\times$ downsampling *vs.* $32\times$ of CNNs), provide a more flexible and granular supervision signal that is less susceptible to domain-specific shifts. Figure 6 provides evidence for this hypothesis, where ConvNeXt-B distilled from all ViTs provides finer semantic clues.

Scaling Performance with Different Dataset Scales. We evaluate ACTok with 1%, 20%, 50%, and 100% ImageNet using the same CLIP, MAE, and DINOv3 ViT-B teachers to distill a ViT-B student. As shown in Figure 7(a), ACTok consistently improves on FGVC with increasing data scale, showing that it remains effective with scaling data.

Comparison on VQA Task. Following RADIO, we assess multimodal transferability by replacing the LLaVA-1.5 [33] vision encoder with our distilled ViT-

Table 4: Ablation on learnable token and modules.

Configuration	$\mathcal{L}^m$	$\mathcal{L}^f$
w/o learnable token	-	0.29
w/o cross-space	0.18	0.35
w/o intra-space	3.56	2.60
ACTok (ours)	0.20	0.23

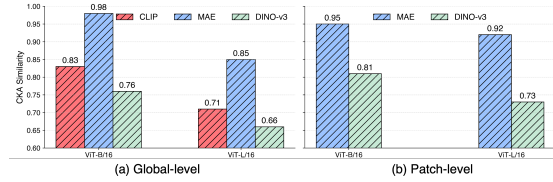


Fig. 8: CKA feature similarity analysis.

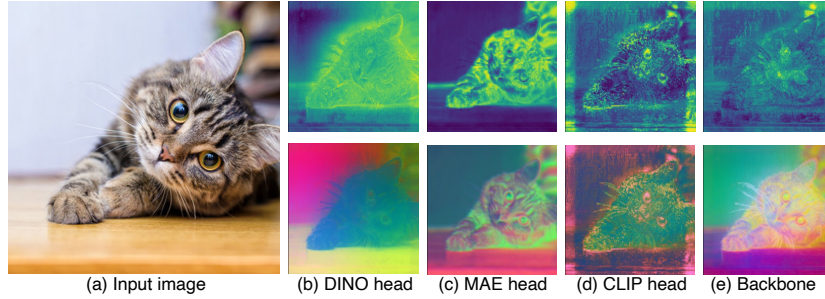


Fig. 9: Cosine similarity and PCA visualization of projected features at different places within the ViT-B/16 student. The input resolution is 2048×2048 , resulting in 128×128 spatial features. **Top:** cosine similarity map between class token and patch tokens. **Bottom:** PCA visualization by projecting features into RGB space.

L under the standard Vicuna-7B-1.5 [66] training protocol. As shown in Figure 7(b), ACTok surpasses E-RADIO-L on POPE and is comparable on VQAv2. RADIOv2.5 lacks these results, while Move-KD uses a different LLaVA setup, so both are excluded.

4.3 Ablation Studies

In this section, we conduct experiments to verify the importance of the introduced learnable token, together with the cross-space and intra-space modules. Besides, we examine the impact of feature alignment objectives as well as the effectiveness of the learning rate scaling strategy in Section 3.4. We also show that the student distilled using the proposed approach can extract robust features against various input resolutions.

Role of the Learnable Token. We distill a ConvNeXt-B student by eliminating the learnable token. The results in Table 4 suggest that the learnable token helps optimize the feature alignment loss. With the interaction between the learnable token and the student’s visual features, the model-agnostic general vision knowledge captured by the learnable token facilitates the alignment between the student and all teachers.

Effectiveness of Cross- and Intra-space Modules. We distill a ConvNeXt-B student by ablating the cross- and intra-space modules to understand their

Table 5: Effect of using different features for classification on the CUB2011 dataset.

Backbone	DINO head	MAE head	CLS head
73.8	77.9 (+4.1)	77.0 (+3.2)	79.7 (+5.9)

Table 6: Impact of different feature alignment objectives.

DINO objective	Aircraft	Caltech101	CUB2011	Average
MSE	89.4	92.8	90.6	90.9
Cosine	92.1	93.8	91.4	92.4

effectiveness. As shown in Table 4, the feature re-organization by the intra-space module is crucial for both learnable token and student feature alignments. Additionally, the cross-space module could help reduce the patch-level alignment loss between students and teachers.

Effect of Alternating Interaction. We conduct fine-grained classification on the CUB2011 dataset, with ResNet-50 as the student backbone. We adopt the learnable token projected to DINO, MAE, and classification heads after alternating interaction modules for classification. From Table 5, backbone features yield unsatisfactory performance, with a noticeable 5.9% accuracy degradation. With the learnable token being projected onto different heads (DINO, MAE, CLS), we observe a consistent accuracy gain. Moreover, we present a PCA visualization analysis in Figure 9, where the projected DINO head perceives better semantic understanding and exhibits powerful detail-capturing capability. Therefore, we consider features projected onto the DINO space for downstream tasks.

Choices of Alignment Objectives. We evaluate the performance of the ViT-L/16 student trained using different DINO objectives on three fine-grained datasets. As shown in Table 6, using a cosine similarity constraint could better transfer the semantic understanding ability to the student.

CKA Analysis. We also adopt the Central Kernel Alignment (CKA) [32] to quantify the representation similarity between the student and all teachers. As shown in Figure 8, the student shares the highest CKA metric with MAE since they are all trained on the same dataset. With a larger training dataset scale, where potential dataset bias becomes more prominent, the CKA similarity between the student and the DINOv3 teacher decreases, resulting in a performance drop compared to the DINOv3 teacher, as shown in Table 1.

Tuning of Alternating Interaction Modules. As shown in Table 7, learning rate scaling for cross- and intra-space modules could significantly boost downstream performance. This underscores the generalization ability of the aggregated knowledge using the proposed ACTok.

Robustness to Input Resolutions. Heinrich *et al.* [24] observe that the student’s feature tends to diverge when input resolution increases to 1024. However, since all teachers in our design are pre-trained at similar resolutions, the ulti-

Table 7: Impact of learning rate scaling of interaction modules.

lr scaling	CUB	Caltech101	Aircraft
w/o	89.2	94.3	90.3
w/	90.8	94.4	91.1

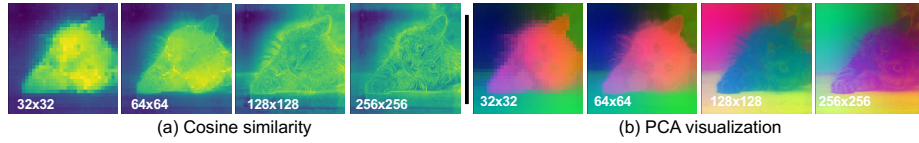


Fig. 10: Robustness to different input resolutions. We visualize the cosine similarity map between cls and patch tokens, and features projected to the RGB space with input image resolutions varying from 512×512 to 4096×4096 .

mate student can extract robust and semantically consistent features when input resolution changes from 512×512 to 4096×4096 as illustrated in Figure 10.

5 Limitations

The primary limitation of ACTok stems from the dataset bias between the teachers’ pre-training source (*e.g.*, DINOv3’s LVD-1689M) and our proxy dataset (ImageNet-1K). While LVD-1689M encompasses diverse web-scale concepts, the ImageNet-1K is predominantly object-centric, which may restrict the transfer of complex multi-object perception capabilities. Furthermore, the strong inductive biases inherent in different architectural families remain a challenge for CNN-based students. Although we demonstrate that an all-ViT teacher ensemble can partially mitigate these effects, more robust mechanisms for combating cross-domain dataset bias in multi-teacher distillation warrant further exploration.

6 Conclusion

In this paper, we present ACTok, a novel framework that consolidates multi-domain knowledge from heterogeneous VFMs via token-mediated aggregation. By introducing learnable tokens as neural proxies and an alternating interaction mechanism, we effectively decouple student learning from the semantic conflicts of divergent teacher architectures. Our evaluations across diverse vision tasks demonstrate that ACTok achieves competitive performance against the strongest VFMs. Crucially, it achieves these results using only ImageNet-1K, proving orders of magnitude more data-efficient than billion-scale paradigms. Furthermore, our empirical findings highlight that homogeneous ViT teacher ensembles optimally alleviate dataset bias for CNN-based students. In summary, ACTok offers an economical, architecture-agnostic, and scalable solution for distilling the collective intelligence of vision foundation models.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62322113, 62376156), as well as the Shanghai Municipal Special Program for Basic Research on General AI Foundation Models (Grant No. 2025SHZDZX025G15).

References

1. Ahn, S., Hu, S.X., Damianou, A., Lawrence, N.D., Dai, Z.: Variational information distillation for knowledge transfer. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 9163–9171 (2019)
2. Asif, U., Tang, J., Harrer, S.: Ensemble knowledge distillation for learning improved and efficient networks. *arXiv preprint arXiv:1909.08097* (2019)
3. Ba, J., Caruana, R.: Do deep nets really need to be deep? In: *Proceedings of the Annual Conference on Neural Information Processing Systems*. pp. 1–9 (2014)
4. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254* (2021)
5. Beyer, L., Zhai, X., Royer, A., Markeeva, L., Anil, R., Kolesnikov, A.: Knowledge distillation: A good teacher is patient and consistent. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 10925–10934 (2022)
6. Cao, J., Zhang, Y., Huang, T., Lu, M., Zhang, Q., An, R., Ma, N., Zhang, S.: Movekd: Knowledge distillation for vlms with mixture of visual encoders. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 19846–19856 (2025)
7. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 9650–9660 (2021)
8. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: *Proceedings of the International Conference on Machine Learning*. pp. 1597–1607 (2020)
9. Cho, J.H., Hariharan, B.: On the efficacy of knowledge distillation. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 4794–4802 (2019)
10. Contributors, M.: Mmsegmentation: Openmmlab semantic segmentation toolbox and benchmark (2020)
11. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. In: *Proceedings of the International Conference on Learning Representations*. pp. 1–11 (2024)
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 248–255 (2009)
13. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pp. 4171–4186 (2019)
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020)
15. Fan, J., Li, C., Liu, X., Yao, A.: Scalekd: Strong vision transformers could be excellent teachers. In: *Proceedings of the Annual Conference on Neural Information Processing Systems*. pp. 63290–63315 (2024)
16. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding* **106**(1), 59–70 (2007)
17. Fukuda, T., Suzuki, M., Kurata, G., Thomas, S., Cui, J., Ramabhadran, B.: Efficient knowledge distillation from an ensemble of teachers. In: *Interspeech*. pp. 3697–3701 (2017)

18. Gadre, S.Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., et al.: Datacomp: In search of the next generation of multimodal datasets. In: *Proceedings of the Annual Conference on Neural Information Processing Systems*. vol. 36, pp. 27092–27112 (2023)
19. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. *International Journal of Computer Vision* **129**(6), 1789–1819 (2021)
20. Hao, Z., Guo, J., Han, K., Tang, Y., Hu, H., Wang, Y., Xu, C.: One-for-all: Bridge the gap between heterogeneous architectures in knowledge distillation. In: *Proceedings of the Annual Conference on Neural Information Processing Systems*. pp. 79570–79582 (2023)
21. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 16000–16009 (2022)
22. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 9729–9738 (2020)
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 770–778 (2016)
24. Heinrich, G., Ranzinger, M., Yin, H., Lu, Y., Kautz, J., Tao, A., Catanzaro, B., Molchanov, P.: Radiov2. 5: Improved baselines for agglomerative vision foundation models. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 22487–22497 (2025)
25. Heo, B., Kim, J., Yun, S., Park, H., Kwak, N., Choi, J.Y.: A comprehensive overhaul of feature distillation. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 1921–1930 (2019)
26. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015)
27. Huang, Z., Wang, N.: Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219* (2017)
28. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: *Proceedings of the International Conference on Machine Learning*. pp. 4904–4916 (2021)
29. Ketkar, N., Moolayil, J.: Convolutional neural networks. In: *Deep learning with Python: learn best practices of deep learning models with PyTorch*, pp. 197–242. Springer (2021)
30. Kim, J., Park, S., Kwak, N.: Paraphrasing complex network: Network compression via factor transfer. In: *Proceedings of the Annual Conference on Neural Information Processing Systems*. pp. 1–10 (2018)
31. 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. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 4015–4026 (2023)
32. Kornblith, S., Norouzi, M., Lee, H., Hinton, G.: Similarity of neural network representations revisited. In: *Proceedings of the International Conference on Machine Learning*. pp. 3519–3529 (2019)
33. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 26296–26306 (2024)
34. Liu, Y., Zhang, W., Wang, J.: Adaptive multi-teacher multi-level knowledge distillation. *Neurocomputing* **415**, 106–113 (2020)

35. Liu, Y., Cao, J., Li, B., Hu, W., Ding, J., Li, L.: Cross-architecture knowledge distillation. In: *Proceedings of the Asian Conference on Computer Vision*. pp. 3396–3411 (2022)
36. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 11976–11986 (2022)
37. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017)
38. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151* (2013)
39. Mirzadeh, S.I., Farajtabar, M., Li, A., Levine, N., Matsukawa, A., Ghasemzadeh, H.: Improved knowledge distillation via teacher assistant. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. pp. 5191–5198 (2020)
40. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*
41. 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 (2023)
42. 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: *Proceedings of the International Conference on Machine Learning*. pp. 8748–8763. *PmLR* (2021)
43. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 12179–12188 (2021)
44. Ranzinger, M., Heinrich, G., Kautz, J., Molchanov, P.: Am-radio: Agglomerative vision foundation model reduce all domains into one. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 12490–12500 (2024)
45. Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J.: Denseclip: Language-guided dense prediction with context-aware prompting. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 18082–18091 (2022)
46. Ren, X., Wang, Z., Hou, L., Tang, P., Wang, G., Ma, C.: Grounding everything in tokens for multimodal large language models. *arXiv preprint arXiv:2512.10554* (2025)
47. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fit-nets: Hints for thin deep nets. In: *Proceedings of the International Conference on Learning Representations*. pp. 1–10 (2015)
48. Sarıyıldız, M.B., Weinzaepfel, P., Lucas, T., De Jorge, P., Larlus, D., Kalantidis, Y.: Dune: Distilling a universal encoder from heterogeneous 2d and 3d teachers. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 30084–30094 (2025)
49. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: *Proceedings of the European Conference on Computer Vision*. pp. 746–760 (2012)

50. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khali-dov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
51. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced trans-former with rotary position embedding. *Neurocomputing* **568**, 127063 (2024)
52. Tian, K., Jiang, Y., Lin, C., Wang, L., Yuan, Z., et al.: Designing bert for convolu-tional networks: Sparse and hierarchical masked modeling. In: *Proceedings of the International Conference on Learning Representations*. pp. 1–13 (2023)
53. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: *Proceedings of the International Conference on Machine Learning*. pp. 10347–10357 (2021)
54. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
55. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 5294–5306 (2025)
56. Wei, Y., Hu, H., Xie, Z., Zhang, Z., Cao, Y., Bao, J., Chen, D., Guo, B.: Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. arXiv preprint arXiv:2205.14141 (2022)
57. Wu, K., Peng, H., Zhou, Z., Xiao, B., Liu, M., Yuan, L., Xuan, H., Valenzuela, M., Chen, X.S., Wang, X., et al.: Tinyclip: Clip distillation via affinity mimicking and weight inheritance. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 21970–21980 (2023)
58. Yang, Z., Shou, L., Gong, M., Lin, W., Jiang, D.: Model compression with two-stage multi-teacher knowledge distillation for web question answering system. In: *Proceedings of the International Conference on Web Search and Data Mining*. pp. 690–698 (2020)
59. You, S., Xu, C., Xu, C., Tao, D.: Learning from multiple teacher networks. In: *Pro-ceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 1285–1294 (2017)
60. Yuan, F., Shou, L., Pei, J., Lin, W., Gong, M., Fu, Y., Jiang, D.: Reinforced multi-teacher selection for knowledge distillation. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. pp. 14284–14291 (2021)
61. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928 (2016)
62. Zhang, C., Han, D., Qiao, Y., Kim, J.U., Bae, S.H., Lee, S., Hong, C.S.: Faster segment anything: Towards lightweight sam for mobile applications. arXiv preprint arXiv:2306.14289 (2023)
63. Zhang, W., Liu, Y., Ran, W., Ma, C.: Cross-architecture distillation made simple with redundancy suppression. arXiv preprint arXiv:2507.21844 (2025)
64. Zhang, Y., Ma, X., Bai, Y., Wang, H., Fu, Y.: Accessing vision foundation mod-els via imagenet-1k. In: *Proceedings of the International Conference on Learning Representations*. pp. 1–14 (2025)
65. Zhao, H., Sun, X., Dong, J., Chen, C., Dong, Z.: Highlight every step: Knowledge distillation via collaborative teaching. *IEEE Transactions on Cybernetics* **52**(4), 2070–2081 (2020)
66. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems* pp. 46595–46623 (2023)

67. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 633–641 (2017)
68. Zhou, C., Li, X., Loy, C.C., Dai, B.: Edgesam: Prompt-in-the-loop distillation for on-device deployment of sam. *arXiv preprint arXiv:2312.06660* (2023)
69. Zhu, X., Gong, S., et al.: Knowledge distillation by on-the-fly native ensemble. In: *Proceedings of the Annual Conference on Neural Information Processing Systems*. pp. 1–11 (2018)
70. Zuchniak, K.: Multi-teacher knowledge distillation as an effective method for compressing ensembles of neural networks. *arXiv preprint arXiv:2302.07215* (2023)