Enabling Visual Foundation Models to Teach Compact Students via Mixture of Distillation

Xinye Yang[1*], **Shang Wang**[2*], **Li Luking** and **Yipeng Chen**[3]
[1]Newcastle University
[2]Independtent Researcher
[3]University of Science and Technology Beijing
c0078451@newcastle.ac.uk, email.shangwang@gmail.com, z2193896726@yahoo.com

Abstract

In this paper, we present a novel Mixture of Distillation (MoD) framework for distilling lightweight student models using Visual Foundation Models (VFMs) as teachers. Knowledge distillation (KD) is a crucial training strategy for improving model performance. However, conventional KD methods face two main challenges: (1) selecting & training appropriate teacher models and (2) designing effective knowledge distillation techniques. To address the first challenge, we leverage recent VFMs like CLIP, Grounding DINO, and SAM as teachers, capitalizing on their remarkable zero-shot generalization abilities and low fine-tuning requirements for new tasks, thereby avoiding expensive retraining of teachers. For the second challenge, our MoD framework focuses on extracting and decomposing the feature and logit knowledge from VFMs into multiple knowledge experts, which capture modality-specific information across batches, channels, and instances. Each knowledge expert undergoes separate projections, reshaping, normalization, and learnable magnitude operations. Then, we employ sparse knowledge gates with a softmax function followed by a KeepTopK operation for different knowledge experts. In this way, our MoD not only bridges the distillation gap between VFMs and students but also allows the adaptive transfer of useful knowledge across different domains. Extensive experiments on various classification, detection, and medical segmentation tasks validate the effectiveness of our approach with other models. Moreover, our MoD framework demonstrates the potential for transferring zero-shot abilities from VFMs without relying on ground-truth labels. Notably, our MoD achieves impressive performance, attaining 72.48% for RepViT with 76.20% CLIP teacher on ImageNet-1K without annotations.

1 Introduction

The development of sizeable deep models [Krizhevsky *et al.*, 2012] in computer vision continues to be significant in recent years. However, it is increasingly apparent that these models often encounter issues such as redundancy and high computational resource requirements. To tackle these challenges, various model compression methods have emerged, aiming to enhance the efficiency and compactness of deep models [Li *et al.*, 2024c; Li *et al.*, 2024e; Dong *et al.*, 2024; Dong *et al.*, 2025b; Li *et al.*, 2024d; **?**; Gu *et al.*, 2025; Li *et al.*, 2025; Dong *et al.*, 2023b; Dong *et al.*, 2025a; Wei *et al.*, 2024]. Among these compression techniques, knowledge distillation has emerged as a highly effective approach [Hinton *et al.*, 2015]. Knowledge distillation involves transferring knowledge from a large, cumbersome model (the teacher model) to a smaller, more streamlined model (the student model). This approach enables the student model to balance efficiency and accuracy, making it suitable for deployment on resource-constrained devices.

Conventional KD methods always focus on extracting useful information from teacher models and reducing teacher-student gaps. For example, different methods are proposed to transfer teachers' knowledge of logits, features, and sample relationships. Other KD methods present smart designs regarding knowledge transformations and distillation losses. However, previous studies rarely address two critical challenges: Selecting and efficiently training teacher models and choosing the suitable teacher knowledge for transfer in different scenarios. (1) Teacher training issue: Although KD methods can reuse existing pre-trained models as teachers under the public dataset, they still have to train teachers models from scratch once they come to new datasets. This additional teacher training process brings large time costs and computational resource consumption. Some self-distillation methods attempt to remove the teacher model via auxiliary structures or losses, but their performance could be better than the teacher-guided methods. (2) Knowledge selection issue: Different knowledge performs differently under various tasks and datasets. For example, logit knowledge performs optimally on the classification task but consistently sub-optimal vs features knowledge for object detection. Simply involving many kinds of knowledge always results in gradient conflicts and optimization difficulties. These difficulties naturally present questions: ***How to efficiently obtain teachers and select the best knowledge for the new scenarios?***

Recently, VFMs, such as CLIP [Radford *et al.*, 2021], DINO [Zhang *et al.*, 2022], and SAM [Kirillov *et al.*, 2023], have reshaped the landscape of computer vision and ma-

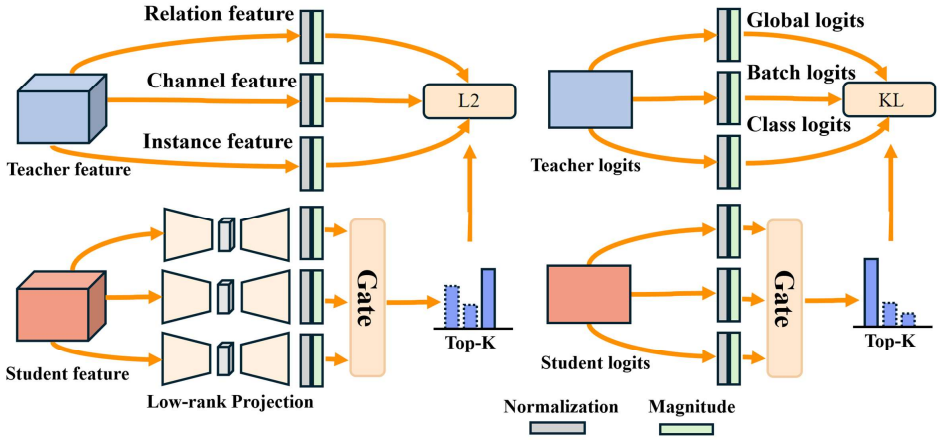

Figure 1: Detailed processes in our MoD framework. First, the feature dimensions between the teacher and student are aligned via low-rank projections, followed by normalization and magnitude scaling operations applied to the knowledge. Sparse knowledge gates are then computed to selectively transfer relevant components based on the student's feature and logit information. The gated knowledge is transferred by calculating distances, using an L2 loss for features and a KL loss for logits.

chine learning with their impressive generalization capabilities across various tasks. These models, pre-trained on vast datasets, can make accurate zero-shot predictions and achieve superior performance after fine-tuning. Advanced parameter-efficient fine-tuning methods, like LoRA [Hu *et al.*, 2021], have further reduced the cost of adapting these VFMs to new tasks, striking a balance between performance and efficiency and making them more accessible and adaptable to a broader range of applications. The strengths of VFMs inspired us to employ them as an alternative to traditional teachers in distillation frameworks. However, these VFMs have larger model scales than traditional teachers, creating more significant teacher-student gaps in architecture and knowledge representations. This highlights the importance of knowledge selection when distilling from VFMs. One key technology in VFMs, the Mixture of Experts (MoE) [Shazeer *et al.*,], provides valuable insights into our knowledge selection designs. MoE enables expert selection for specific tasks via gate sparsity, offering a promising direction for selectively distilling knowledge from VFMs.

Based on the above observations, we propose a novel Mixture of Distillation (MoD) framework that capitalizes on the strengths of VFMs while addressing the challenges posed by their scale and complexity, opening new doors for more accessible applications of these powerful models in distillation. As shown in Figure 1, we first fine-tune partial parameters in the VFM teacher model on the downstream datasets. In this way, we can quickly obtain ultra-powerful teachers with preferable feature encoders and minimal training cost compared to traditional teacher training. Then, we decompose the feature and logit knowledge from VFMs into multiple knowledge components. For feature knowledge, we align dimensions using separate low-rank projections and transform the aligned features into relation, instance, and channel feature knowledge through reshaping, normalization, and learnable magnitude operations. Similarly, we derive instance, relation, and class logit knowledge by reshaping, normalizing, and re-weighting the original logits. Our MoD framework employs sparse knowledge gates to adaptively transfer helpful knowledge to the student model tailored to the specific domain or task. Our sparse gates use a softmax function followed by a KeepTopK operation to introduce sparsity in the gate selection process. This knowledge selection strategy reduces knowledge conflicts and optimization difficulties arising from multiple losses in the distillation process. Finally, the selected knowledge serves as an input in the pairwise distance function, acting as an optimization objective for knowledge distillation. In this way, our MoD bridges the gap between large-scale VFMs and compact student models, enabling efficient knowledge transfer and task-specific adaptation.

We conduct a comprehensive evaluation to assess the effectiveness of our MoD framework across diverse tasks, datasets, and student architectures. When applied to classification tasks, MoD consistently outperforms other distillation methods, such as KD, DIST, and MGD, in both annotation-free and annotation-based scenarios. Notably, our MoD achieves an accuracy 72.48 for RepViT and 72.85 for ViT-Small on ImageNet without ground truth labels. Moving on to object detection tasks, MoD demonstrates remarkable performance improvements compared to student baselines, FGD, and PKD across various architectures, encompassing two-stage, one-stage, anchor-based, and anchor-free detectors. These improvements are observed for both baseline and stronger backbone models. For instance, MoD achieves an impressive 47.7 AP on RetinaNet, surpassing FGD and PKD by significant margins. In the realm of medical image analysis, MoD enhances the performance of student models such as TinySAM and EfficientSAM when Medical-SAM is used as the teacher model. These findings provide strong evidence for the efficacy and applicability of MoD, demonstrating its ability to consistently improve performance across diverse computer vision tasks. In summary, the key contributions of the proposed approach are:

- We propose a novel Mixture of Distillation (MoD) frame-

work that capitalizes on the strengths of VFMs while addressing the challenges posed by their scale and complexity, opening new doors for more accessible applications of these powerful models in distillation.

- Our MoD framework employs sparse knowledge gates to adaptively transfer useful knowledge to the student model, tailored to the specific domain or task, reducing knowledge conflicts and optimization difficulties arising from multiple losses in the distillation process.
- The comprehensive evaluation of MoD across a wide range of classification, object detection, and medical image analysis tasks demonstrates its effectiveness, robustness, and versatility in consistently improving the performance of compact student models.

2 Related Works

Knowledge distillation. Knowledge Distillation (KD) [Bucila *et al.*, 2006] has seen the development of various approaches and techniques for training a student model using the knowledge from a high-capacity teacher model. Pioneer studies, such as [Bucila *et al.*, 2006; Hinton *et al.*, 2015], utilized soft logit outputs from teachers to provide additional supervision during training. Recent logits KD methods present varying dynamic temperature factors [Li *et al.*, 2023b] or target class non-target class decoupling [Zhao *et al.*, 2022] to reduce the teacher-student gap [Li and Jin, 2022]. Besides logits KD, another type is feature KD methods [Romero *et al.*, 2015; Li, 2022] focusing on transferring rich feature knowledge within teacher models. These methods present many knowledge transformation designs and distance functions to align teacher-student features in shape and semantics. Other methods like RKD [Park *et al.*, 2019] explore the structural information and contrastive learning methods like SSKD [Xu and others, 2020] incorporates self-supervised learning into the distillation process. Moreover, customized distillation methods and design strategies are based on specific tasks. For example, detection distillation methods focus on transferring knowledge in instance and foreground features. While these existing methods suggest superior distillation solutions, they often overlook the teacher and knowledge selection processes, limiting their practical applications. In contrast, our approach efficiently fine-tunes VFMs as teachers and incorporates MoD for selective knowledge transfer, harnessing the potential of these powerful models. Unlike distillation tuning methods [Li *et al.*, 2023a; Li *et al.*, 2024b; Sun *et al.*, 2024a; Dong *et al.*, 2023a; Li *et al.*, 2024a] that require additional hyperparameter optimization overhead, our MoD method does not introduce an additional stage. By efficiently fine-tuning VFMs and selectively distilling relevant knowledge, our approach bridges the gap between these powerful and compact student models.

Visual foundation model. Vision foundation models (VFMs) trained on extensive datasets represent a significant advancement in computer vision, serving as the cornerstone for a wide range of downstream tasks. CLIP [Radford *et al.*, 2021], a multimodal model, learns joint representations of images and text and has shown remarkable success in zero-shot classification tasks by aligning the visual and textual domains through contrastive learning. Derivative works of CLIP have expanded its applications to various tasks [Ali and Khan, 2023]. For object detection, DINO [Zhang *et al.*, 2022] builds upon the DETR [Dai *et al.*, 2021] architecture and utilizes self-supervised learning to allow the model to learn rich feature representations without extensive manual annotation. Grounding DINO [Liu *et al.*, 2023], a variant of DINO, combines it with grounded pre-training for open-set object detection. In the segmentation domain, SAM [Kirillov *et al.*, 2023] is a prompt-based model that learns to generate segmentation masks based on user-provided prompts, with training involving a large dataset and a unique approach. Recently, techniques like Parameter-Efficient Fine-Tuning (PEFT) [Zhang *et al.*, 2023] and Mixture-of-Experts (MoE) [Zoph *et al.*, 2022] have been proposed to improve the efficiency of VFMs on different tasks. PEFT adds Adaptation Layers or employs Low-Rank Adaptation [Liu *et al.*, 2024] to fine-tune a minimal subset of model weights, enabling efficient adaptation of large-scale models. MoE, by combining specialized experts and a flexible gating mechanism, enhances the performance, efficiency, and adaptability of VFMs. While some methods (*e.g.*, TinyCLIP [Wu *et al.*, 2023]) attempt to distill large VFMs into small VFMs, many compact models still require many parameters and components for various tasks. Additionally, the training process for distilling small VFMs can be expensive. In contrast, our approach focuses on enhancing the training of lightweight models in traditional scenarios, allowing for a broader range of applications.

3 Methodology: Mixture of Distillation

3.1 Reviewing Knowledge Distillation

Knowledge distillation transfers knowledge from a large, powerful teacher model T to a compact student network S. Specifically, the goal is to transfer the teacher's feature and logit knowledge, denoted as f_T and p_T respectively, to the student's outputs f_S and p_S. This is typically achieved by minimizing a distance function between the transformed teacher and student outputs:

$$\mathcal{L}_{KD} = \mathcal{D}_f\big(\mathcal{T}_f^S(f^S), \mathcal{T}_f^T(f^T)\big) + \mathcal{D}_p\big(\mathcal{T}_p(p^S/\tau), \mathcal{T}_p(p^T/\tau)\big), \tag{1}$$

where $\mathcal{T}_f$ and $\mathcal{T}_p$ are transformations applied to the features and logits, respectively, $\mathcal{D}_f(\cdot,\cdot)$ and $\mathcal{D}_p(\cdot,\cdot)$ are distance functions measuring the difference between feature representations and logits, and τ is a temperature scaling factor.

3.2 Knowledge Decomposition and Representation

For our MoD framework in Figure 1, we decompose the knowledge from VFMs into multiple knowledge components. This decomposition process aims to effectively capture and represent the rich knowledge encoded in the VFMs, enabling efficient knowledge transfer to the student model.

Feature knowledge representation. Let $f \in \mathbb{R}^{B\times C\times H\times W}$ denote the feature maps extracted from the VFM teacher model, where B, C, H, and W represent the batch size, number of channels, height, and width, respectively. We first align the feature dimensions of teacher-student models using separate low-rank projections: $Pro(f) = \mathbf{W}_1\mathbf{W}_2 f$, where $\mathbf{W}_1$ and $\mathbf{W}_2$ are learnable projection matrices which are added to

the student model and updated during the distillation process. To decompose the teacher's feature knowledge, we reshape the aligned feature f into three components: relation feature $f_r \in \mathbb{R}^{B\times(CHW)}$, channel feature $f_c \in \mathbb{R}^{C\times(NHW)}$, and instance feature $f_i \in \mathbb{R}^{B\times C\times(H/n)\times(W/n)}$. Here, the relation feature captures the spatial relationships between feature elements, the channel feature encodes the channel-wise information, and the instance feature represents the localized instance-level features. Finally, we apply a normalization operation $||\cdot||_{norm}$ and scale the normalized features with a learnable magnitude parameter m as:

$$\mathcal{T}^S_{f_{\{r,c,i\}}} = m_{\{r,c,i\}} \times ||Pro_{\{r,c,i\}}(f^S_{\{r,c,i\}})||_{norm_{\{r,c,i\}}}, \quad (2)$$

$$\mathcal{T}^T_{f_{\{r,c,i\}}} = m_{\{r,c,i\}} \times ||f^T_{\{r,c,i\}}||_{norm_{\{r,c,i\}}}, \quad (3)$$

where $\mathcal{T}^S_{f_{r,c,i}}$ and $\mathcal{T}^T_{f_{r,c,i}}$ represent the transformed feature knowledge components of the student and teacher, respectively. The magnitude parameters $m_{r,c,i}$ allow the framework to adaptively scale the importance of different knowledge components. For each transformed teacher-student feature pair, we employ the L2 distance as the distillation loss:

$$\mathcal{L}_{\{r,c,i\}} = \mathcal{D}_{L2}\big(\mathcal{T}^S_{f\ \{r,c,i\}}(f^S_{\{r,c,i\}}), \mathcal{T}^T_{f\ \{r,c,i\}}(f^T_{\{r,c,i\}})\big), \quad (4)$$

where $\mathcal{D}_{L2}(\cdot,\cdot)$ is the L2 distance function, encouraging the student's transformed features to match those of the teacher.

Logits knowledge representation. Similar to feature knowledge, we derive instance, relation, and class logit knowledge from the original logits $p \in \mathbb{R}^{N\times C}$ of the VFM teacher, where N is the number of instances, and C is the number of classes. The global logit knowledge, denoted as $p_g \in \mathbb{R}^{N\times C}$, is the original logit matrix. The batch logit knowledge captures the pairwise relationships between instance logits, denoted as $p_b \in \mathbb{R}^{N\times N}$. The class logit knowledge, denoted as $p_c \in \mathbb{R}^{C\times C}$, captures the logit patterns across different classes. We apply normalization and re-weighting operations to these logit knowledge components:

$$\mathcal{T}_{p_{\{g,b,c\}}} = m_{\{g,b,c\}} \times ||p_{\{g,b,c\}}||_{norm_{\{g,b,c\}}}, \quad (5)$$

where $m_{g,b,c}$ is a learnable scalar weight parameter for the respective logit knowledge component.

For the logits part of the knowledge, we select the Kullback-Leibler (KL) divergence as the distillation loss for each transformed teacher-student logit pair:

$$\mathcal{L}_{\{g,b,c\}} = \mathcal{D}_{KL}\left(\mathcal{T}_{p_{\{g,b,c\}}}(p^S_{\{g,b,c\}}/\tau), \mathcal{T}_{p_{\{g,b,c\}}}(p^T_{\{g,b,c\}}/\tau)\right), \quad (6)$$

where $\mathcal{D}_{KL}(\cdot,\cdot)$ is the KL divergence function, and τ is a temperature scaling factor.

3.3 Knowledge Selection with Sparse Gating

While decomposing the VFM knowledge into multiple components is beneficial, not all components may be equally relevant or useful for a given task or scenario. Transferring irrelevant or conflicting knowledge can hinder the student model's performance and lead to optimization difficulties. To address this, our MoD framework employs sparse knowledge gates $\mathbf{G}$ to selectively transfer the most relevant knowledge components to the student model while suppressing irrelevant or conflicting knowledge. The sparse knowledge gates $\mathbf{G}$ are computed as follows:

$$\mathbf{G}_f = KeepTopK(Softmax(\mathbf{W}_f \times f_s, k)), \quad (7)$$

$$\mathbf{G}_p = KeepTopK(Softmax(\mathbf{W}_p \times f_p, k)), \quad (8)$$

where $\mathbf{W}_f$ and $\mathbf{W}_p$ are learnable parameters, Softmax$(\cdot)$ is the softmax function applied across the knowledge component dimension, and $KeepTopK(\cdot, k)$ is an operation that keeps the k largest values in the input vector and sets the remaining values to zero, introducing sparsity. The sparse gate matrices $\mathbf{G}_f$ and $\mathbf{G}_p$ are then used to select the relevant features and logits Knowledge for each instance:

$$\mathcal{L}_{\{r,c,i\}} = \mathcal{D}_{L2}\big(\mathbf{G}^{r,c,i}_f \times \mathcal{T}^S_{f\ \{r,c,i\}}(f^S_{\{r,c,i\}}), \mathcal{T}^T_{f\ \{r,c,i\}}(f^T_{\{r,c,i\}})\big), \quad (9)$$

$$\mathcal{L}_{\{g,b,c\}} = \mathcal{D}_{KL}\left(\mathbf{G}^{g,r,c}_p \times \mathcal{T}_{p_{\{g,b,c\}}}(p^S_{\{g,b,c\}}), \mathcal{T}_{p_{\{g,b,c\}}}(p^T_{\{g,b,c\}})\right), \quad (10)$$

where $\mathbf{G}^{r,c,i}_f$ and $\mathbf{G}^{g,r,c}_p$ are the sparse gates for the respective feature and logit knowledge components. The sparse gating mechanism introduces sparsity in the knowledge selection process, allowing the MoD framework to adaptively transfer only the most relevant knowledge components to the student model. This selective knowledge transfer reduces optimization conflicts and enables more efficient learning by focusing on the most useful information for the given task.

3.4 Optimization Objective

The final optimization objective for our MoD framework is a combination of the pairwise distance between the selected knowledge components of the teacher and student models, as well as the original task loss of the student:

$$\mathcal{L}_{\text{MoD}} = \mathcal{L}_{original} + \mathcal{L}_r + \mathcal{L}_c + \mathcal{L}_i + \mathcal{L}_g + \mathcal{L}_b + \mathcal{L}_c. \quad (11)$$

where $\mathcal{L}_{original}$ denotes the original task loss with ground-truth labels. In particular, on the classification task, CLIP teacher models produce quite high-quality predictions. Thus, we use directly CLIP models' predictions as an alternative to human labels to teach student. Such annotation-free distillation also bring very promising results, detailed in the following experiments.

4 Experiments

4.1 Experiments on Image Classification

Implementation Details. We assess our framework on fine-grained classification datasets (*e.g.*, Stanford Cars[Krause *et al.*, 2013], Oxford Pets[Parkhi *et al.*, 2012], CIFAR-100[Alex, 2009] and Food-101[Bossard *et al.*, 2014]), large-scale dataset ImageNet-1K[Deng *et al.*, 2009]. We use annotation-free as a teacher and fine-tune it in an annotation-based setting and zero-shot prediction in an annotation-free setting. We perform the same distillation and data augmentation settings for all comparison methods. Detailed implementations are in the Appendix.

Comparison results on fine-grained datasets. We conduct comparative experiments on multiple datasets, using the CLIP ViT-L/14 model as the teacher and EfficientNet-B1 and RepViT M1.1 as the student models. The experimental

Teacher	CIFAR100		Caltech101		Food101		StanfordCars		OxfordPets	
Annotation-free										
CLIP ViT-L/14	77.50		92.80		93.80		78.80		93.50	
Student	ENet	RepViT	ENet	RepViT	ENet	RepViT	ENet	RepViT	ENet	RepViT
KD [Hinton *et al.*, 2015]	74.61	74.81	81.16	81.05	84.08	87.47	67.75	73.52	80.92	83.35
DIST [Huang *et al.*, 2022]	74.49	72.65	80.47	80.07	84.50	85.28	73.50	73.70	82.31	80.10
CAT-KD[Guo *et al.*, 2023]	74.01	73.51	80.21	77.73	84.88	87.14	68.47	72.24	83.74	82.19
FreeKD [Zhang *et al.*, 2024]	75.43	73.25	79.57	79.52	86.49	86.62	70.55	73.42	75.67	76.90
MGD [Yang *et al.*, 2022]	72.05	72.25	80.47	79.90	83.42	86.76	72.70	72.45	77.33	80.27
MoD (ours)	**75.72**	**76.15**	**82.56**	**83.51**	**87.03**	**87.51**	**75.32**	**75.13**	**84.62**	**84.75**
Annotation-based										
CLIP ViT-L/14	87.50		96.00		95.90		90.50		95.10	
Student	ENet	RepViT	ENet	RepViT	ENet	RepViT	ENet	RepViT	ENet	RepViT
KD [Hinton *et al.*, 2015]	78.20	78.56	93.30	93.60	87.60	88.40	76.80	78.20	86.30	86.90
DIST [Huang *et al.*, 2022]	78.55	78.59	93.56	93.98	87.30	88.50	77.10	77.92	86.22	86.40
CAT-KD[Guo *et al.*, 2023]	78.05	78.54	92.64	93.27	87.02	87.72	75.62	77.62	85.98	86.17
FreeKD [Zhang *et al.*, 2024]	78.42	78.67	92.72	94.28	88.12	88.07	76.72	78.18	86.17	86.47
MGD [Yang *et al.*, 2022]	78.62	78.68	94.35	94.42	87.20	88.90	77.70	78.46	86.65	87.25
MoD (ours)	**79.82**	**80.25**	**95.25**	**95.50**	**89.80**	**90.15**	**78.85**	**79.22**	**88.20**	**88.35**

Table 1: Top-1 accuracies (%) of different distillation methods. We adopt a ViT-L/14 model from CLIP [Radford *et al.*, 2021] as the teacher network. For the student models, we select three efficient yet compact models: EfficientNet-B1 [Tan and Le, 2019] (ENet) and RepVit M1.1 [Wang *et al.*, 2023] (RepViT).

Teacher: CLIP ViT-L/14 → Student: RepViT M1.1		
Method	*Annotation-free*	*Annotation-based*
Teacher	76.20	85.40
Student	NA	79.40
KD [Hinton *et al.*, 2015]	70.75	80.70
DIST [Huang *et al.*, 2022]	71.03	80.85
CAT-KD[Guo *et al.*, 2023]	71.29	81.03
ReviewKD [Pengguang Chen and Jia, 2021]	70.88	80.65
CWD [Shu *et al.*, 2021]	71.10	81.02
MGD [Yang *et al.*, 2022]	71.25	81.25
CTKD [Li *et al.*, 2023c]	70.86	80.95
FreeKD [Zhang *et al.*, 2024]	71.45	81.55
SDD [Luo, 2024]	71.28	81.32
Logit Stand. [Sun *et al.*, 2024b]	71.13	81.08
MoD (ours)	**72.48**	**82.65**

Teacher: CLIP ViT-L/14 → Student: ViT-Small		
Method	*Annotation-free*	*Annotation-based*
Teacher	76.20	85.40
Student	NA	79.90
KD [Hinton *et al.*, 2015]	71.15	81.20
DIST [Huang *et al.*, 2022]	71.32	81.05
CAT-KD[Guo *et al.*, 2023]	71.26	81.18
ReviewKD [Pengguang Chen and Jia, 2021]	71.22	81.25
CWD [Shu *et al.*, 2021]	71.41	81.40
MGD [Yang *et al.*, 2022]	71.62	81.60
CTKD [Li *et al.*, 2023c]	71.23	81.36
FreeKD [Zhang *et al.*, 2024]	71.67	81.90
SDD [Luo, 2024]	71.56	81.98
Logit Stand. [Sun *et al.*, 2024b]	71.34	81.67
MoD (ours)	**72.85**	**83.05**

Table 2: Top-1 classification accuracy (%) results on ImageNet validation set.

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Two-stage detectors						
T: Grounding-DINO-L	60.3	77.6	66.4	45.1	64.4	75.7
S: Faster RCNN-R50	38.4	59.0	42.0	21.5	42.1	50.3
FGD [Yang *et al.*, 2021]	38.0	58.3	41.3	21.2	41.8	49.7
PKD [Cao *et al.*, 2022]	36.5	57.5	39.2	19.6	40.2	50.5
MoD (ours)	40.5	63.2	44.2	24.5	45.0	51.7
One-stage detectors						
T: Grounding-DINO-L	60.3	77.6	66.4	45.1	64.4	75.7
S: RetinaNet-R50	37.4	56.7	39.6	20.0	40.7	49.7
FGD [Yang *et al.*, 2021]	36.8	56.2	38.7	19.4	40.2	49.1
PKD [Cao *et al.*, 2022]	37.2	56.6	39.3	19.6	40.5	49.1
MoD (ours)	40.2	61.0	43.1	23.6	43.7	54.1
Anchor-free detectors						
T: Grounding-DINO-L	60.3	77.6	66.4	45.1	64.4	75.7
S: RepPoints-R50	38.6	59.6	41.6	22.5	42.2	50.4
FGD [Yang *et al.*, 2021]	38.1	59.3	41.4	21.8	41.6	49.8
PKD [Cao *et al.*, 2022]	37.1	59.2	41.3	22.1	41.8	50.1
MoD (ours)	41.7	63.6	45.1	24.8	45.3	55.0
Anchor-based detectors						
T: Grounding-DINO-L	60.3	77.6	66.4	45.1	64.4	75.7
S: ATSS-R50	39.4	57.6	42.8	23.6	42.9	50.3
FGD [Yang *et al.*, 2021]	38.3	56.6	41.8	22.1	41.3	49.6
PKD [Cao *et al.*, 2022]	37.9	56.2	41.5	22.7	41.7	48.9
MoD (ours)	41.2	60.6	45.0	24.7	45.0	53.1

Table 3: **Object detection performance with baseline settings on COCO val set.** T: teacher. S: student.

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Two-stage detectors						
T: Grounding-DINO-L	60.3	77.6	66.4	45.1	64.4	75.7
S: Faster RCNN-X-101	41.2	62.1	45.1	24.0	45.5	53.5
FGD [Yang *et al.*, 2021]	40.9	61.5	44.3	23.3	45.6	53.3
PKD [Cao *et al.*, 2022]	40.3	61.8	44.4	23.8	44.8	52.7
MoD (ours)	43.0	66.0	47.6	27.0	47.8	54.3
One-stage detectors						
T: Grounding-DINO-L	60.3	77.6	66.4	45.1	64.4	75.7
S: RetinaNet-PVTv2 (B4)	46.3	67.0	49.6	29.0	50.1	62.7
FGD [Yang *et al.*, 2021]	45.8	66.7	49.2	28.1	49.6	62.5
PKD [Cao *et al.*, 2022]	46.2	66.8	59.4	28.7	50.0	62.3
MoD (ours)	47.7	70.4	51.6	33.3	51.7	61.9
Anchor-free detectors						
T: Grounding-DINO-L	60.3	77.6	66.4	45.1	64.4	75.7
S: RepPoints-X-101 (DCN)	44.2	65.5	47.8	26.2	48.4	58.5
FGD [Yang *et al.*, 2021]	43.5	64.7	47.2	25.1	47.9	57.5
PKD [Cao *et al.*, 2022]	43.8	65.2	47.3	25.7	48.1	58.2
MoD (ours)	46.9	69.7	51.2	32.2	51.5	60.9
Anchor-based detectors						
T: Grounding-DINO-L	60.3	77.6	66.4	45.1	64.4	75.7
S: ATSS-R101	41.5	59.9	45.2	24.2	45.9	53.3
FGD [Yang *et al.*, 2021]	40.6	59.2	44.7	23.5	45.2	52.4
PKD [Cao *et al.*, 2022]	40.2	58.7	43.6	23.2	44.8	52.3
MoD (ours)	43.1	62.7	47.0	25.9	47.2	56.0

Table 4: **Object detection performance with stronger backbones on COCO val set.**T: teacher. S: student.

Dataset	CVC-ClinicDB		Kvasir		Isic2018		Synapse		ACDC	
Method	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice
Medical-SAM (Teacher)	89.7	94.3	88.9	93.6	85.7	92.1	81.5	87.4	86.4	89.8
TinySAM (Student)	87.0	91.9	84.9	90.5	81.5	88.2	76.8	84.0	80.2	84.5
TinySAM (MoD)	88.9	93.9	85.5	92.2	82.8	89.8	78.4	85.5	81.1	85.8
EfficientSAM (Student)	85.5	91.3	85.3	90.9	81.7	88.6	77.5	84.5	82.2	87.9
EfficientSAM (MoD)	88.5	93.1	87.6	92.7	84.6	91.1	79.9	85.9	84.9	89.0

Table 5: Quantitative evaluation results on 2D medical datasets (CVC-ClinicDB, and Kvasir and ISIC2018) and two 3D medical datasets (Synapse and ACDC).

Ablation		Knowledge Representation					Knowledge Selection				
GT	Inputs	MLP	LoRA	S-LoRA	Norm	Magnitude	All	Top-2 (Random)	Top-1(Random)	Top-2 (MoD)	Top-1 (MoD)
w	Feature	78.65	78.80	79.15	79.20	79.47	77.30	78.18	78.24	79.54	79.61
wo	Feature	74.58	74.67	74.91	75.36	75.47	73.11	73.67	73.86	75.37	75.47
w	Logits	NA	NA	NA	79.30	79.45	77.10	78.19	78.24	79.60	79.64
wo	Logits	NA	NA	NA	75.33	75.49	73.87	73.96	74.07	75.48	75.58
w	F+L	79.05	79.18	79.32	79.48	79.59	77.68	78.56	78.52	79.75	79.82
wo	F+L	74.98	75.05	75.26	75.48	75.61	73.58	74.08	74.25	75.65	75.72

Table 6: Ablation study of various components in MoD for EfficientNet-B1 on CIFAR100 dataset.

results in Table 1 provide a comprehensive analysis of our approach, MoD, compared to other distillation methods in two scenarios: Annotation-free and Annotation-based. In the annotation-free setting, where no labeled data is available, our MoD approach consistently outperforms other methods across all datasets and student models. For instance, on the CIFAR100 dataset, MoD achieves accuracies of 75.72% and 76.15% for EfficientNet-B1 and RepViT M1.1, respectively, outperforming the second-best method, KD, by 1.11% and 1.34%, respectively. Similarly, on the Caltech101 dataset, MoD attains accuracies of 82.56% and 83.51% for the two student models, surpassing other method by a significant margin. Moving to the annotation-based scenario, MoD continues to exhibit superiority over other distillation methods. These gains are particularly significant compared to KD, DIST, and MGD, indicating its ability to effectively distill knowledge from the teacher model using labeled data and additional annotations. In an Annotation-free or Annotation-based scenario, MoD consistently improves the performance of the student models compared to other distillation methods. These findings suggest that MoD can be a valuable tool for model compression and knowledge transfer in various domains and applications.

Comparison results on large-scale datasets. The results on the ImageNet-1K[Deng *et al.*, 2009] dataset, as shown in Table **??**. In the annotation-free setting, our MoD approach consistently outperforms other methods by significant margins. For the RepViT M1.1 student model, MoD achieves a 1.23% gain than MGD. Similarly, MoD attains 72.85% accuracy on ViT-Small, outperforming other methods. These results demonstrate the effectiveness of our approach in distilling knowledge from the teacher model without relying on any labeled data. Our MoD approach maintains its superior performance in the annotation-based setting with labeled data for training. For the RepViT M1.1 student model, MoD achieves 82.65% accuracy on RepViT and 83.05% on ViT-Small, surpassing other methods with more than 1.45% gains. The consistent performance gains of our MoD approach across both annotation-free and annotation-based settings, showcase its robustness and versatility.

4.2 Experiments on Object Detection

Implementation Details. For detection on MS-COCO dataset [Caesar *et al.*, 2018], we conduct experiments with baseline and stronger backbone settings on two-stage, one-stage, anchor-free, anchor-based detectors. We fine-tune Grounding-DINO-L on the MS-COCO and use it as the teacher detector. We follow the same distillation setups in previous methods [Yang *et al.*, 2021]. Detailed implementations are in the Appendix.

Comparison results. Table 3 and Table 4 demonstrate the remarkable gains achieved by our method with Grounding-DINO-L as teacher across various object detection settings. In the baseline setting (see Table 3), our method consistently outperforms the student baseline and other methods (*e.g.*, FGD and PKD) by substantial margins. For two-stage detectors like Faster R-CNN, our method achieves a 40.5 AP, surpassing baseline and FGD by a considerable 2.1 and 2.5 AP, respectively. Similar trends are observed for one-stage detectors (RetinaNet), anchor-free detectors (RepPoints), and anchor-based detectors (ATSS), where our method outperforms the student baseline by 2.8, 3.1, and 1.8 in AP, respectively, and consistently outperforms FGD and PKD across all metrics. When employing stronger backbones (see Table 4), such as ResNeXt-101 and PVTv2-B4, our method maintains its superiority, further widening the performance gap compared to the student baseline and other distillation methods. For example, with RetinaNet-PVTv2 (B4), our method achieves an impressive 47.7 AP of 47.7, outperforming FGD (45.8) and PKD (46.2). This significant gain is also observed for anchor-free and anchor-based detectors, where our method outperforms the student baselines by 1.4, 2.7, and 1.6 AP, respectively.

4.3 Experiments on Medical Image Segmentation

Implementation Details. we conduct experiments on a diverse set of medical datasets. This ongoing evaluation includes three 2D medical datasets [Fang *et al.*, 2022], ISIC2018, CVC-ClinicDB, and Kvasir, and two 3D medical datasets [Wang *et al.*, 2021], Synapse and ACDC. Our experiments employ

CIFAR100	Caltech101	Food101	ImageNet	Faster RCNN	RetinaNet	Kvasir	Synapse
f_r, p_g	f_c, p_g	f_i, p_r	f_i, p_r	f_c, p_g	f_i, p_g	f_i, p_c	f_c, p_c

Table 7: Knowledge selection results in our MoD on different tasks.

	Teacher(CLIP)			Loss-weights				Temperature			
MoD	R50	ViT-B/32	ViT-L/14	0.1	0.5	1	5	1	4	8	16
	78.95	79.47	79.82	79.15	79.37	79.82	79.65	79.25	79.82	79.75	79.42

Table 8: Various distillation configuration in MoD for EfficientNet-B1 on CIFAR100 dataset.

SAM-based models [Ma and Wang, 2023] with one-point prompts. Detailed implementations are in the Appendix.

Comparison results. Table 5 presented the performance of two student models, TinySAM with TinyVit backbone and EfficientSAM with Efficientformer backbone, with and without our MoD, using the Medical-SAM model [Ma and Wang, 2023] as the teacher. Focusing on the 2D medical datasets, we can observe that the MoD framework consistently improves the performance of both student models. Similarly, the EfficientSAM model exhibits notable performance gains when trained with MoD, surpassing the baseline by 3.0% and 1.8%, respectively. The effectiveness of our MoD framework is further demonstrated by the 3D medical datasets Synapse and ACDC. For example, on the Synapse dataset, the TinySAM model with MoD achieves 78.4% mIoU, and the EfficientSAM model with MoD attains 2.4%$\sim$ 1.4% gains, respectively. These results demonstrate the ability of our MoD to effectively distill knowledge from the Medical-SAM teacher model on 2D and 3D segmentation.

4.4 Ablation Study

Knowledge Representation. Table 6 reports the various designs for MoD. We observe that the magnitude operation consistently achieves the highest performance across all settings. For instance, with GT and features as inputs, Magnitude attains an accuracy of 79.59%, outperforming MLP baseline by 0.54%. However, when solely relying on logits as inputs, Magnitude and Norm exhibit comparable performance, with Magnitude slightly edging out Norm by 0.15% when GT is present and 0.16 percentage points without GT. The normalize and separate LoRA-based projection (S-LoRA) also demonstrate robust performance gains, trailing slightly behind magnitude operation.

Knowledge Selection. Table 6 compares our sparse gate-based selection to random selection. We find that our methods in Top-1 and Top-2 selection consistently outperform the random selection approaches. This suggests that our knowledge selection strategies, which prioritize the most important elements, are more effective at capturing and retaining the relevant knowledge representations for the given tasks, leading to higher performance in both the with ground-truth and without ground-truth settings. The gains are particularly notable for the logits-only scenarios, underscoring the importance of our targeted selection approach in optimizing the knowledge representation when working with limited input information.

Selected Knowledge. Table 7 presents a summary of selected knowledge from various experiments. Notably, it can be observed that across different downstream tasks, instance and channel features, as well as global knowledge, consistently emerge as important factors. In classification tasks with many categories, relational features and class logits play a dominant role in the selection process.

Sensitivity Analysis of Configurations. Table 8 presents ablation study on KD configurations, we find the following: (1) The ViT-L/14 teacher achieves the highest performance at 79.82%, followed by ViT-B/32 at 79.47% and ResNet50 at 78.95%. (2) Loss weight of 1 produces the best results at 79.82%, while lower or higher weights lead to slightly reduced performance. (3) Temperature of 4 yields the highest accuracy of 79.82%, indicating this level of logit scaling is most effective for knowledge distillation. These findings highlight the importance of carefully selecting the teacher model, optimizing the loss-weight configuration, and tuning the temperature scaling to achieve the best performance with the MoD.

5 Conclusion

In this paper, we propose a MoD framework to transfer knowledge from large-scale VFMs like CLIP, DINO effectively, and SAM to compact student models. The MoD framework bridges the gap between large-scale VFMs and compact student models, enabling efficient knowledge transfer and task-specific adaptation. Extensive evaluations show that MoD consistently outperforms other distillation methods across classification, object detection, and medical imaging tasks while transferring VFM zero-shot abilities to students in annotation-free settings. We sincerely hope that our methods will provide insights to the community and enhance the accessibility of KD in practical applications.

Contribution Statement

Corresponding author: Yipeng Chen. Xinye Yang* and Shang Wang* have equal contributions (* denotes equal contributions).

References

[Alex, 2009] Krizhevsky Alex. Learning multiple layers of features from tiny images. *https://www. cs. toronto. edu/kriz/learning-features-2009-TR. pdf*, 2009.

[Ali and Khan, 2023] Muhammad Ali and Salman Khan. Clip-decoder: Zeroshot multilabel classification using multimodal clip aligned representations. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision, pages 4675–4679, 2023.

[Bossard *et al.*, 2014] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, 2014.

[Bucila *et al.*, 2006] Cristian Bucila, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *KDD*, 2006.

[Caesar *et al.*, 2018] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *CVPR*, pages 1209–1218, 2018.

[Cao *et al.*, 2022] Weihan Cao, Yifan Zhang, Jianfei Gao, Anda Cheng, Ke Cheng, and Jian Cheng. Pkd: General distillation framework for object detectors via pearson correlation coefficient. *NeurIPS*, 35:15394–15406, 2022.

[Dai *et al.*, 2021] Xiyang Dai, Yinpeng Chen, Jianwei Yang, Pengchuan Zhang, Lu Yuan, and Lei Zhang. Dynamic detr: End-to-end object detection with dynamic attention. In *ICCV*, 2021.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. 2009.

[Dong *et al.*, 2023a] Peijie Dong, Lujun Li, and Zimian Wei. Diswot: Student architecture search for distillation without training. In *CVPR*, 2023.

[Dong *et al.*, 2023b] Peijie Dong, Lujun Li, Zimian Wei, Xin Niu, Zhiliang Tian, and Hengyue Pan. Emq: Evolving training-free proxies for automated mixed precision quantization. *arXiv preprint arXiv:2307.10554*, 2023.

[Dong *et al.*, 2024] Peijie Dong, Lujun Li, Zhenheng Tang, Xiang Liu, Xinglin Pan, Qiang Wang, and Xiaowen Chu. Pruner-zero: Evolving symbolic pruning metric from scratch for large language models. In *ICML*, 2024.

[Dong *et al.*, 2025a] Peijie Dong, Lujun Li, Zhenheng Tang, Xiang Liu, Zimian Wei, Qiang Wang, and Xiaowen Chu. Parzc: Parametric zero-cost proxies for efficient nas. In *AAAI*, 2025.

[Dong *et al.*, 2025b] Peijie Dong, Lujun Li, Yuedong Zhong, Dayou Du, Ruibo Fan, Yuhan Chen, Zhenheng Tang, Qiang Wang, Wei Xue, Yike Guo, et al. Stbllm: Breaking the 1-bit barrier with structured binary llms. In *ICLR*, 2025.

[Fang *et al.*, 2022] Huihui Fang, Fei Li, Huazhu Fu, Xu Sun, Xingxing Cao, Jaemin Son, Shuang Yu, Menglu Zhang, Chenglang Yuan, Cheng Bian, et al. Refuge2 challenge: Treasure for multi-domain learning in glaucoma assessment. *arXiv preprint arXiv:2202.08994*, 2022.

[Gu *et al.*, 2025] Hao Gu, Wei Li, Lujun Li, Zhu Qiyuan, Mark Lee, Shengjie Sun, Wei Xue, and Yike Guo. Delta decompression for moe-based llms compression. In *ICML*, 2025.

[Guo *et al.*, 2023] Ziyao Guo, Haonan Yan, Hui Li, and Xiaodong Lin. Class attention transfer based knowledge distillation. In *CVPR*, 2023.

[Hinton *et al.*, 2015] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[Hu *et al.*, 2021] J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021.

[Huang *et al.*, 2022] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. *arXiv preprint arXiv:2205.10536*, 2022.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[Krause *et al.*, 2013] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV Workshops*, pages 554–561, 2013.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[Li and Jin, 2022] Lujun Li and Zhe Jin. Shadow knowledge distillation: Bridging offline and online knowledge transfer. In *NeuIPS*, 2022.

[Li *et al.*, 2023a] Lujun Li, Peijie Dong, Zimian Wei, and Ya Yang. Automated knowledge distillation via monte carlo tree search. In *ICCV*, 2023.

[Li *et al.*, 2023b] Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. Curriculum temperature for knowledge distillation. In *AAAI*, 2023.

[Li *et al.*, 2023c] Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. Curriculum temperature for knowledge distillation. In *AAAI*, 2023.

[Li *et al.*, 2024a] Lujun Li, Yufan Bao, Peijie Dong, Chuanguang Yang, Anggeng Li, Wenhan Luo, Qifeng Liu, Wei Xue, and Yike Guo. Detkds: Knowledge distillation search for object detectors. In *ICML*, 2024.

[Li *et al.*, 2024b] Lujun Li, Peijie Dong, Anggeng Li, Zimian Wei, and Ya Yang. Kd-zero: Evolving knowledge distiller for any teacher-student pairs. *NeuIPS*, 2024.

[Li *et al.*, 2024c] Lujun Li, Peijie, Zhenheng Tang, Xiang Liu, Qiang Wang, Wenhan Luo, Wei Xue, Qifeng Liu, Xiaowen Chu, and Yike Guo. Discovering sparsity allocation for layer-wise pruning of large language models. In *NeuIPS*, 2024.

[Li *et al.*, 2024d] Lujun Li, Haosen Sun, Shiwen Li, Peijie Dong, Wenhan Luo, Wei Xue, Qifeng Liu, and Yike Guo. Auto-gas: Automated proxy discovery for training-free generative architecture search. ECCV, 2024.

[Li *et al.*, 2024e] Wei Li, Lujun Li, Mark Lee, and Shengjie Sun. Als: Adaptive layer sparsity for large language models via activation correlation assessment. In *NeuIPS*, 2024.

[Li *et al.*, 2025] Wei Li, Lujun Li, You-Liang Huang, Mark G. Lee, Shengjie Sun, Wei Xue, and Yike Guo. Structured mixture-of-experts LLMs compression via singular value decomposition. In *ICML*, 2025.

[Li, 2022] Lujun Li. Self-regulated feature learning via teacher-free feature distillation. In *ECCV*, 2022.

[Liu *et al.*, 2023] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[Liu *et al.*, 2024] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation, 2024.

[Luo, 2024] Shicai Wei Chunbo Luo Yang Luo. Scale decoupled distillation. *arXiv preprint arXiv:2403.13512*, 2024.

[Ma and Wang, 2023] Jun Ma and Bo Wang. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023.

[Park *et al.*, 2019] Wonpyo Park, Yan Lu, Minsu Cho, and Dongju Kim. Relational knowledge distillation. In *CVPR*, 2019.

[Parkhi *et al.*, 2012] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012.

[Pengguang Chen and Jia, 2021] Hengshuang Zhao Pengguang Chen, Shu Liu and Jiaya Jia. Distilling knowledge via knowledge review. In *CVPR*, 2021.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[Romero *et al.*, 2015] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *ICLR*, 2015.

[Shazeer *et al.*,] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR*.

[Shu *et al.*, 2021] Changyong Shu, Yifan Liu, Jianfei Gao, Lin Xu, and Chunhua Shen. Channel-wise distillation for semantic segmentation. In *ICCV*, 2021.

[Sun *et al.*, 2024a] Haosen Sun, Lujun Li, Peijie Dong, Zimian Wei, and Shitong Shao. Auto-das: Automated proxy discovery for training-free distillation-aware architecture search. ECCV, 2024.

[Sun *et al.*, 2024b] Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. Logit standardization in knowledge distillation. In *CVPR*, 2024.

[Tan and Le, 2019] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114. PMLR, 2019.

[Wang *et al.*, 2021] Wenxuan Wang, Chen Chen, Meng Ding, Hong Yu, Sen Zha, and Jiangyun Li. Transbts: Multimodal brain tumor segmentation using transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 109–119. Springer, 2021.

[Wang *et al.*, 2023] Ao Wang, Hui Chen, Zijia Lin, Hengjun Pu, and Guiguang Ding. Repvit: Revisiting mobile cnn from vit perspective. *arXiv preprint arXiv:2307.09283*, 2023.

[Wei *et al.*, 2024] Zimian Wei, Peijie Dong, Zheng Hui, Anggeng Li, Lujun Li, Menglong Lu, Hengyue Pan, and Dongsheng Li. Auto-prox: Training-free vision transformer architecture search via automatic proxy discovery. In *AAAI*, 2024.

[Wu *et al.*, 2023] Kan Wu, Houwen Peng, Zhenghong Zhou, Bin Xiao, Mengchen Liu, Lu Yuan, Hong Xuan, Michael Valenzuela, Xi Stephen Chen, Xinggang Wang, et al. Tinyclip: Clip distillation via affinity mimicking and weight inheritance. In *ICCV*, 2023.

[Xu and others, 2020] Guodong Xu et al. Knowledge distillation meets self-supervision. In *ECCV*, pages 588–604. Springer, 2020.

[Yang *et al.*, 2021] Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and global knowledge distillation for detectors. *arXiv preprint arXiv:2111.11837*, 2021.

[Yang *et al.*, 2022] Zhendong Yang, Zhe Li, Mingqi Shao, Dachuan Shi, Zehuan Yuan, and Chun Yuan. Masked generative distillation. *arXiv preprint arXiv:2205.01529*, 2022.

[Zhang *et al.*, 2022] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

[Zhang *et al.*, 2023] Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning, 2023.

[Zhang *et al.*, 2024] Yuan Zhang, Tao Huang, Jiaming Liu, Tao Jiang, Kuan Cheng, and Shanghang Zhang. Freekd: Knowledge distillation via semantic frequency prompt. In *CVPR*, 2024.

[Zhao *et al.*, 2022] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *CVPR*, pages 11953–11962, 2022.

[Zoph *et al.*, 2022] Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. St-moe: Designing stable and transferable sparse expert models, 2022.