SWISS ARMY KNIFE: SYNERGIZING BIASES IN KNOWLEDGE FROM VISION FOUNDATION MODELS FOR MULTI-TASK LEARNING

Yuxiang Lu[1]* **Shengcao Cao**[2]* **Yu-Xiong Wang**[2]
[1]Shanghai Jiao Tong University [2]University of Illinois Urbana-Champaign
luyuxiang_2018@sjtu.edu.cn {cao44,yxw}@illinois.edu

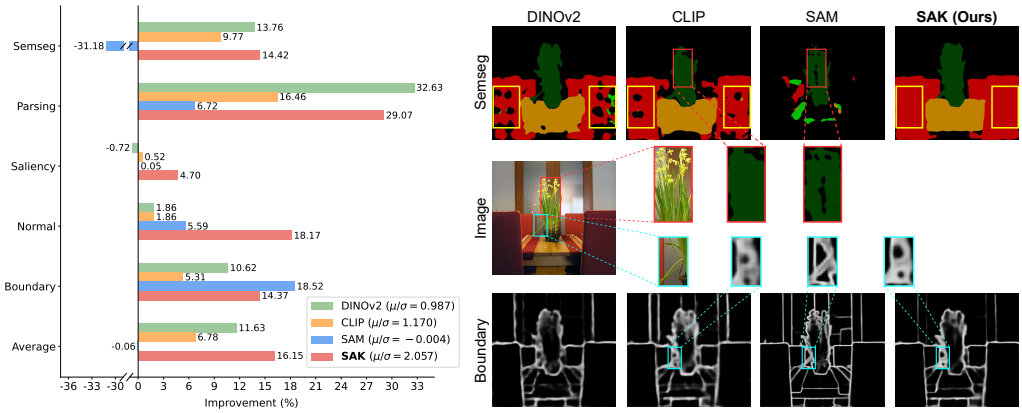

Figure 1: *(Left)* **Quantitative analysis of representation biases** in Vision Foundation Models (VFMs), including DINOv2, CLIP, and SAM, on the PASCAL-Context dataset across five vision tasks, all using the ViT-B backbones with pretrained parameters frozen. VFMs exhibit advantages and disadvantages across different downstream tasks when compared to a conventional ImageNet-pretrained backbone. Our SAK model, distilled from these VFM teachers, achieves the *best average performance with more balanced improvements*, as indicated by its larger ratio of mean improvement to standard deviation (μ/σ). *(Right)* **Qualitative comparison of representation biases** through representative examples from semantic segmentation and boundary detection tasks. DINOv2 captures localized features but occasionally confuses semantic categories; CLIP excels in object-level understanding but lacks fine pixel-level details; SAM produces precise masks in both tasks due to higher input resolution but struggles with semantic knowledge. Our SAK successfully combines the *precise boundary detection* of SAM with the *accurate semantic understanding* of DINOv2 and CLIP. Further details are discussed in Section 2.

ABSTRACT

Vision Foundation Models (VFMs) have demonstrated outstanding performance on numerous downstream tasks. However, due to their inherent representation biases originating from different training paradigms, VFMs exhibit advantages and disadvantages across distinct vision tasks. Although amalgamating the strengths of multiple VFMs for downstream tasks is an intuitive strategy, effectively exploiting these biases remains a significant challenge. In this paper, we propose a novel and versatile "Swiss Army Knife" (SAK) solution, which adaptively distills knowledge from a committee of VFMs to enhance multi-task learning. Unlike existing methods that use a single backbone for knowledge transfer, our approach preserves the unique representation bias of each teacher by collaborating the lightweight Teacher-Specific Adapter Path modules with the Teacher-Agnostic Stem. Through dynamic selection and combination of representations with Mixture-of-Representations Routers, our SAK is capable of synergizing the complementary strengths of multiple VFMs. Extensive experiments show that our SAK remarkably outperforms prior state of the arts in multi-task learning by 10% on the NYUD-v2 benchmark, while also providing a flexible and robust framework that can readily accommodate more advanced model designs. Project page: https://innovator-zero.github.io/SAK/.

*Equal Contribution

1 INTRODUCTION

Vision Foundation Models (VFMs), such as DINOv2 (Oquab et al., 2024), CLIP (Radford et al., 2021), and SAM (Kirillov et al., 2023), have gained significant attention due to their impressive performance on various downstream tasks. This underscores the importance of integrating VFMs into Multi-Task Learning (MTL) (Caruana, 1997; Zhang & Yang, 2021; Yu et al., 2024), which aims to jointly learn multiple tasks with a single network, thereby enhancing model efficiency and performance, with broad applications in areas like autonomous driving (Ishihara et al., 2021).

In computer vision, multi-task models typically use a shared encoder to extract general features for all tasks, as they share a common interpretation of visual input (Ye & Xu, 2023a). A straightforward approach is to directly replace the encoder backbone with a VFM. However, VFMs are pretrained on diverse datasets, image resolutions, and objectives, introducing *representation biases* when applied as feature extractors for downstream tasks. Our empirical study in Figure 1 reveals that these inherent biases yield both advantages and disadvantages across different tasks, with *no single model achieving consistently superior performance across all domains*. These findings highlight the challenge of accomplishing comprehensive improvements in MTL using VFMs, pointing to the demand for collaborative utilization of multiple VFMs to exploit their complementary strengths.

Several existing works attempt an intuitive solution by extracting image features through multiple VFMs and then concatenating or fusing these features for later decoding (Lin et al., 2023; Kar et al., 2024; Zong et al., 2024; Tong et al., 2024a;b; Man et al., 2024). While this enhances visual encoding, it comes with a major drawback: The inference of all vision encoders drastically increases computational costs, along with the memory and storage requirements due to the large-scale parameters, rendering it less practical for real-world applications.

Therefore, recent works (Ranzinger et al., 2024b; Shang et al., 2024; Sariyildiz et al., 2024) propose more efficient frameworks by distilling multiple VFM teachers into a single student model, which can deliver competitive results on downstream benchmarks. Despite the progress, this many-to-one distillation risks eliminating the representation biases of the VFM teachers, potentially limiting the model's ability to capitalize on their individual strengths for specific tasks. Zong et al. (2024) further point out that biased information from VFMs can lead to performance degradation when naively fused. Moreover, when matching one student to multiple teachers, reconciling diverse biases in shared parameters could induce optimization conflicts. Our pilot study in Table 2 shows that the student trained by many-to-one distillation does not consistently surpass the teachers in their respective proficient tasks.

To overcome these limitations, we propose a novel approach named **SAK**, with the goal to build a **S**wiss **A**rmy **K**nife model from a committee of VFMs to synergize their complementary strengths and enhance performance across multiple downstream tasks. Considering the key challenge of *preserving the representation biases* while ensuring model efficiency for deployment, we introduce a multi-teacher knowledge distillation framework. This framework incorporates a shared Teacher-Agnostic Stem alongside multiple Teacher-Specific Adapter Path modules, which produce specialized representations aligned with each corresponding VFM teacher. During distillation, the Teacher-Agnostic Stem is optimized simultaneously by gradients from all VFM teachers, thereby capturing universal knowledge. Meanwhile, the Teacher-Specific Adapter Paths accommodate the heterogeneous representation biases of each teacher, explicitly learning their diverse model characteristics.

Building on the reproduction of representation biases, the next step is to amalgamate the committee's expertise by *exploiting the individual biases*. Specifically, we treat each group of representations as a knowledgeable expert and design a Mixture-of-Representations Router. This router dynamically weighs and combines the most relevant representations, bridging the gap between general-purpose knowledge and task-specific characteristics. The collaboration of these modules allows the student to harness both commonalities and differences of the teachers, facilitating smoother and more comprehensive knowledge transfer. Furthermore, SAK is a highly flexible framework that can further benefit from more advanced architectural designs (*e.g.*, stronger task-specific decoders) and more powerful models (*e.g.*, larger teachers), offering a general solution to multi-task visual learning.

Our contributions are summarized as follows:

Table 1: **Comparison of Vision Foundation Models.** Although all utilize the same Vision Transformer (ViT) backbone, they greatly differ in their training paradigms, including data, image resolutions, and training objectives, which lead to diverse representation biases.

Model	Training Dataset	Dataset Size	Resolution	Objective
ViT (Dosovitskiy et al., 2021)	ImageNet-1k/21k	1.2M/14.2M	384	Supervised classification
DINOv2 (Oquab et al., 2024)	LVD-142M	142M	518	Discriminative self-supervised learning
CLIP (Radford et al., 2021)	WebImageText	400M	224	Image-text contrastive learning
OpenCLIP (Cherti et al., 2023)	LAION-2B	2B	384	Image-text contrastive learning
SAM (Kirillov et al., 2023)	SA-1B	11M+1B	1024	Supervised promptable segmentation

- We systematically analyze the distinct representation biases of Vision Foundation Models, which result in varying advantages and disadvantages across tasks, underscoring the importance of preserving these biases during distillation from multiple VFM teachers.
- We propose SAK, an efficient and effective solution that distills knowledge from VFM teachers into a Teacher-Agnostic Stem with Teacher-Specific Adapter Path modules, sharing common knowledge while retaining the biases. We also introduce Mixture-of-Representations Routers to adaptively amalgamate the complementary and specialized strengths for downstream tasks.
- We evaluate SAK on two widely-used multi-task benchmarks, PASCAL-Context and NYUD-v2, showing it remarkably outperforms previous multi-teacher VFM distillation methods and state-of-the-art multi-task models in both performance and robustness.
- SAK offers high flexibility and scalability, supporting a broad variety of VFM teachers and downstream tasks, and is compatible with various adapter, router, or decoder head architectures.

2 Representation Biases in Vision Foundation Models

In this section, we investigate the representation biases of Vision Foundation Models on multiple downstream tasks through empirical studies. We select three representative state-of-the-art VFMs: (1) *DINOv2* (Oquab et al., 2024), which claims to excel in dense prediction tasks such as semantic segmentation and depth estimation; (2) *CLIP* (Radford et al., 2021) and its reproduction, *OpenCLIP* (Cherti et al., 2023), which are recognized for capturing language-aligned semantics and employed as vision encoders in vision-language models; and (3) *SAM* (Kirillov et al., 2023), which achieves outstanding performance in promptable segmentation. For CLIP and SAM, we use only their image encoders for representation learning.

As summarized in Table 1, although all these VFMs utilize Vision Transformers (ViT) (Dosovitskiy et al., 2021) as backbones, they differ significantly in their training paradigms regarding datasets, dataset sizes, image resolutions, and training objectives. Consequently, *the representations learned by these models embed heterogeneous biases*, causing each model to focus on different aspects of image features and exhibit strengths and weaknesses in specific tasks.

We conduct comprehensive quantitative and qualitative experiments using the three VFMs on five dense prediction tasks from the PASCAL-Context dataset (Mottaghi et al., 2014). Among these tasks, intuitively, *semantic segmentation* and *human parsing* require high-level semantics of objects and localized features to generate accurate masks. *Saliency detection* demands an overall understanding of the image to identify its main contents, while *surface normal estimation* and *object boundary detection* depend more on fine-grained representations for precise predictions.

We further provide a pilot study to show the inferiority of ignoring representation biases in knowledge distillation from multiple VFM teachers, which validates the significance of addressing this problem and motivates the development of our methodology.

2.1 Quantitative Analysis

To quantitatively analyze the representation biases, we evaluate the performance of the three VFMs directly transferred to each downstream task. We first freeze the models to generate image representations based on their pretrained knowledge, and then train a decoder head to produce final predictions for each task. DINOv2 and CLIP operate at a resolution of 512 on the downstream dataset, while SAM uses an input size of 1,024 as required by its pipeline. All feature maps are resized to $1/4$ of the output resolution before being passed to the head. To quantify the advantages of VFMs over the conventional ImageNet-pretrained ViT backbone, we calculate their relative improvement over ViT for each task.

Table 2: **Comparison of a student model trained by many-to-one distillation without preserving representation biases and the oracle derived from VFM teachers.** The oracle selects the best result from the three teachers for each task. The student's 2.34% average underperformance demonstrates the critical importance of maintaining these biases during distillation.

Model	Semseg mIoU↑	Parsing mIoU↑	Saliency maxF↑	Normal mErr↓	Boundary odsF↑
Oracle of teachers	81.18 (DINOv2)	74.38 (DINOv2)	81.48 (CLIP)	16.21 (SAM)	75.89 (SAM)
Student w/o biases	80.18 (↓ 1.23%)	69.13 (↓ 7.06%)	82.72 (↑ 1.52%)	16.00 (↑ 1.30%)	71.16(↓ 6.23%)

Figure 1*(Left)* illustrates how the representation biases in VFMs manifest in varying strengths and weaknesses in different downstream tasks. Specifically, DINOv2 shows significant improvements in two segmentation tasks, particularly excelling in human parsing with a performance gain of over 30%. It also performs well in object boundary detection, benefiting from its strongly localized features learned from the combination of image-level contrastive objective (Caron et al., 2021) and patch-level reconstructive objective iBOT (Zhou et al., 2022). While CLIP achieves lower accuracy than DINOv2 in these three tasks, it still exceeds the baseline by a notable margin of over 5%. Despite being pretrained on a segmentation task, SAM surprisingly underperforms ViT in semantic segmentation, showing a 30% drop, because of its limited semantic understanding—SAM considers solely the object masks and ignores their semantic labels in its promptable segmentation task. However, SAM is the best in surface normal estimation and object boundary detection, exhibiting strength in capturing pixel-level details and object edges. We also compute the ratio of mean improvement μ to standard deviation σ across tasks, which can measure the consistency of improvements. A higher ratio indicates better outcomes, as it reflects larger average improvements with smaller dispersion. We can observe that while DINOv2 demonstrates stronger average enhancement, CLIP attains more balanced results, whereas SAM is inferior in both perspectives.

2.2 Qualitative Analysis

To validate our quantitative findings, we visualize the final predictions for semantic segmentation and boundary detection using an example image in Figure 1*(Right)*. We observe that DINOv2, while being effective at capturing localized features, is less effective than CLIP in semantic perception, as illustrated in the yellow box of Semseg results. In this case, DINOv2 confuses a chair with a sofa, resulting in misclassification as the background (the holes). Although CLIP excels in object-level understanding with its rich semantic knowledge from the language domain, it falls short in generating fine-grained pixel-level masks. This shortcoming arises because CLIP's training objective prioritizes image-level contents that are represented only by the class token, which possibly accounts for its lower performance than DINOv2 in the quantitative analysis.

On the other hand, SAM produces exceptional details in both tasks due to its high input resolution, as demonstrated in the red and cyan boxes. In the red box, a complex scene shows a foreground flower blending into the background, yet SAM accurately detects and labels the background in the segmentation mask. Notably, the background is not annotated in the ground truth, as such precise masking requires significant time and effort. We regard SAM's high resolution as its representation bias, as it stems directly from the model's training paradigm. However, SAM's limitation lies in its semantic knowledge, particularly when integrating semantics from multiple objects. This makes it difficult to attain high-quality semantic segmentation results, even with highly precise masks, echoing the quantitative analysis. We provide additional analysis and discussions in Appendix A.

2.3 Importance of Preserving Representation Biases

In summary, the inherent representation biases in VFMs result in their uneven performance across tasks, *with no single model achieving the best results in all areas*, as reflected in our quantitative analysis. This motivates the idea of combining multiple VFMs to achieve optimal performance in all tasks. Existing methods (Ranzinger et al., 2024b; Shang et al., 2024; Sariyildiz et al., 2024) propose a solution by distilling multiple VFMs into a single student model. However, given that the student model is shared by several teachers, an important question naturally arises: *Should we respect their individual representation biases when combining diverse VFMs?*

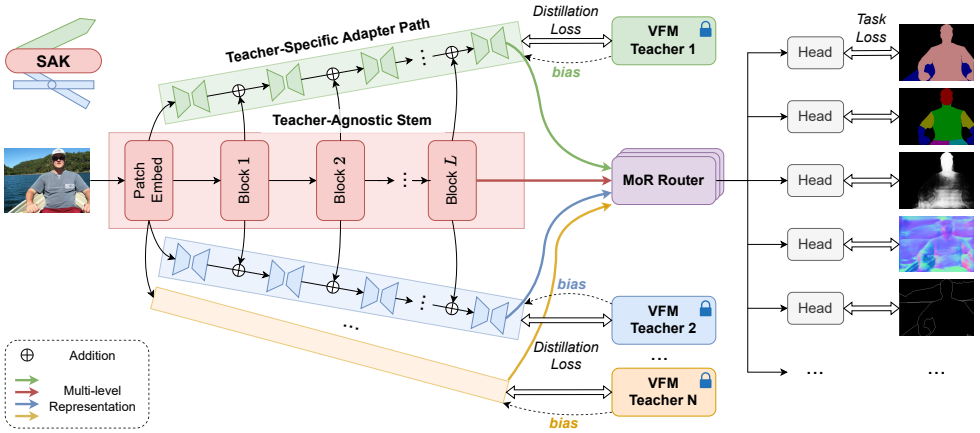

Figure 2: **Overview of our proposed SAK framework,** which distills foundational knowledge from a committee of frozen VFM teachers into an efficient student model. The student model operates like a Swiss Army Knife, with the Teacher-Agnostic Stem (TAS) serving as the main branch to learn universal knowledge among teachers. Each Teacher-Specific Adapter Path (TSAP) acts as a specialized tool to preserve the inherent representation bias of each teacher. Task-specific Mixture-of-Representations (MoR) Routers are then employed to synergize the complementary strengths of the teachers' biases, adaptively combining multi-level representations from both TAS and TSAP to generate tailored features for each task.

To answer this question, we conduct a pilot study where a student model is distilled from three aforementioned VFMs, using only linear aligners to match the student's features with those of the teachers, following the setup of the state-of-the-art method (Ranzinger et al., 2024b). In this approach, the representation biases are not explicitly preserved during distillation, leading the student to learn a unified representation aimed at simultaneously matching all three teachers. We then evaluate its performance on downstream tasks with the same settings as in quantitative analysis in Section 2.1. We compare the student without biases to an oracle derived from the teachers by selecting the best-performing teacher for each task, which represents the optimal performance of teachers.

From Table 2, the answer is clearly **YES**. While the distilled student surpasses the oracle in Saliency and Normal, somewhat validating the effectiveness of prior methods, it suffers from drastic performance degradation in Parsing and Boundary, with a drop of over 6%. Given that both the teachers and student utilize a ViT-B backbone in this study, the performance gap would likely widen with larger models. This demonstrates the limitation of naively transferring knowledge from multiple teachers into a student and highlights the importance of preserving the individual biases, leading us to the key question: ***Can we preserve the representation biases of multiple VFMs during distillation to maximize multi-task performance?*** Our methodology provides a positive answer to this challenge in the following sections.

3 Methodology

3.1 Overview

The overall framework of the proposed SAK is depicted in Figure 2. As a multi-teacher distillation approach, it employs a committee of VFM teachers, including DINOv2, CLIP, and SAM. The student model comprises a **Teacher-Agnostic Stem (TAS)** and multiple **Teacher-Specific Adapter Path (TSAP)** modules. TAS produces general representations shared across all branches, while each TSAP adapts the common representations to align with the specialized domain of its corresponding teacher via distillation. In this approach, the TSAP modules are optimized explicitly to replicate the unique representation biases of the teachers, all in a parameter- and computationally-efficient manner. The resulting feature sets are then passed through task-specific **Mixture-of-Representations (MoR) Routers** for adaptive combination and are finally processed by prediction heads to generate outputs for multiple tasks. We utilize multi-level representations for both the distillation and task decoding procedures, an essential aspect for dense prediction tasks (Ye & Xu, 2022).

3.2 TEACHER-AGNOSTIC STEM & TEACHER-SPECIFIC ADAPTER PATH

We adopt an off-the-shelf Vision Transformer (ViT) (Dosovitskiy et al., 2021) as the Teacher-Agnostic Stem (TAS) and design a lightweight network branch called Teacher-Specific Adapter Path (TSAP) parallel to the main stem. Given a TAS with L blocks, the forward pass for an input image $\boldsymbol{X}$ is expressed as:

$$\boldsymbol{Z}_0 = \texttt{PatchEmbed}(\boldsymbol{X}); \quad \boldsymbol{Z}_l = b_l(\boldsymbol{Z}_{l-1}), \quad l \in \{1, 2, \ldots, L\}, \tag{1}$$

where b_l represents the l-th block, and $\boldsymbol{Z}_l \in \mathbb{R}^{n \times d}$ denotes its intermediate outputs with n tokens of dimension d. Each TSAP module consists of $L+1$ adapters $\{a_l\}, l \in \{0, 1, \ldots, L\}$, with one adapter parallel to each patch embedding layer or transformer block. These adapters process intermediate features to adapt them to the teacher-specific representations $\boldsymbol{R}_l \in \mathbb{R}^{n \times d}$ in a residual manner:

$$\boldsymbol{R}_0 = a_0(\boldsymbol{Z}_0); \quad \boldsymbol{R}_l = a_l(\boldsymbol{R}_{l-1} + \boldsymbol{Z}_l), \quad l \in \{1, 2, \ldots, L\}. \tag{2}$$

We utilize the standard adapter structure (Houlsby et al., 2019), which includes a down-projection layer $\mathbf{W}_{\text{down}} \in \mathbb{R}^{d \times r}$, a GELU non-linearity (Hendrycks & Gimpel, 2016), and an up-projection layer $\mathbf{W}_{\text{up}} \in \mathbb{R}^{r \times d}$, where $r \ll d$ is the reduced dimension. As in prior works (Chen et al., 2022; Mercea et al., 2024), we integrate a learnable scaling factor α and a residual connection from the input $\boldsymbol{R}_{\text{in}} \in \mathbb{R}^{n \times d}$ to the output $\boldsymbol{R}_{\text{out}} \in \mathbb{R}^{n \times d}$, which can be formulated as:

$$\boldsymbol{R}_{\text{out}} = \alpha \texttt{GELU}(\boldsymbol{R}_{\text{in}} \mathbf{W}_{\text{down}}) \mathbf{W}_{\text{up}} + \boldsymbol{R}_{\text{in}}. \tag{3}$$

For a committee of N VFM teachers, we assign a TSAP module with adapters $\{a_l^i\}, l \in \{0, 1, \ldots, L\}$ to the i-th teacher. We then select four evenly distributed blocks from its outputs $\{\boldsymbol{R}_l^i\}$ to form multi-level representations $\{\boldsymbol{R}^i\}_s = \{\boldsymbol{R}_l^i\}, l \in \mathbb{L}_s = \{L/4, L/2, 3L/4, L\}$. Similarly, we have shared multi-level representations $\{\boldsymbol{Z}\}_s$ from TAS. Benefiting from the lightweight adapters, our distilled student model maintains original efficiency, as each TSAP module accounting for less than 5% of the TAS parameters. Consequently, our SAK framework is able to preserve the representation biases from the teachers without significant increases in computational cost, memory usage, or storage, as shown in detailed discussions in Appendix D.1.

3.3 MIXTURE-OF-REPRESENTATIONS ROUTER

As depicted in Figure 2, we treat the representations from TAS $\{\boldsymbol{Z}\}_s$ as a shared expert providing common knowledge, while the representations from each TSAP $\{\boldsymbol{R}^i\}_s, i \in \{1, 2, \ldots, N\}$ serve as proxy experts of VFMs with representation biases mirroring the teachers, resulting in a total of $N + 1$ experts. To optimize the multi-task performance, we leverage the Mixture-of-Experts (MoE) mechanism (Jacobs et al., 1991), which adaptively produces task-specific features from this pool of general-purpose and specialized representations.

To facilitate this, task-specific router networks are trained to generate gate scores for each expert representation, which serve as the weights for a linear combination. As shown in Figure 6, the representations from different VFMs exhibit substantial variation in norm magnitudes. Thus, applying different weights for individual patches within an image can be less effective, as it may disturb the inherent patterns. To address this, we design a Mixture-of-Representations (MoR) Router, which differs from prior works by generating a global gating score across all patches.

Specifically, for each selected level $l \in \mathbb{L}_s$ and downstream task $t \in \mathbb{T}$, our MoR Router r_l^t takes the teacher-agnostic representation $\boldsymbol{Z}_l \in \mathbb{R}^{n \times d}$ as input, projects its channel dimension to $N + 1$ through a two-layer MLP, and then averages over n patches to get a feature vector $\boldsymbol{h}_l^t \in \mathbb{R}^{N+1}$. To improve stability, we incorporate the noisy gating technique (Shazeer et al., 2017) by generating a noise vector $\boldsymbol{e}_l^t \in \mathbb{R}^{N+1}$ through an additional MLP. Then we compute the gating score $\boldsymbol{g}_l^t \in \mathbb{R}^{N+1}$:

$$\boldsymbol{g}_l^t = \texttt{Softmax}(\boldsymbol{h}_l^t + \mathcal{N}(0, 1)\texttt{Softplus}(\boldsymbol{e}_l^t)). \tag{4}$$

The output gating scores are used to calculate the weighted sum of representations at each output level. The fused features are then passed through task-specific heads for final predictions.

3.4 TRAINING PARADIGM

Our training paradigm contains two stages, with teacher parameters always frozen. In the first stage, we train the student model on the ImageNet-1k dataset (Deng et al., 2009; Russakovsky et al.,

2015), focusing on aligning the outputs of the TSAP modules with their respective VFM teachers. ImageNet is chosen due to its diverse and extensive image samples, providing a strong basis for effective knowledge transfer. To maintain fairness—given that conventional ViT backbones are also pretrained on ImageNet—we opt not to use other larger datasets like those utilized in VFMs and RADIO (Ranzinger et al., 2024b). Following previous findings (Ranzinger et al., 2024b; Shang et al., 2024), we employ a combination of cosine distance and smooth-L1 losses for distillation. Let $\boldsymbol{T}_l^i$ be the i-th teacher's representation at a selected level $l \in \mathbb{L}_s$, the overall distillation loss is:

$$\mathcal{L}_{\text{distill}}(\boldsymbol{X}) = \sum_{l \in \mathbb{L}_s} \sum_{i=1}^{N} \left(\alpha \mathcal{L}_{\cos}(\boldsymbol{R}_l^i, \boldsymbol{T}_l^i) + \beta \mathcal{L}_{\text{smooth-L1}}(\boldsymbol{R}_l^i, \boldsymbol{T}_l^i) \right). \tag{5}$$

where $\alpha = 0.9$ and $\beta = 0.1$ are weighting coefficients.

In the second stage, we continue training on the downstream multi-task datasets. The distillation loss is still included, allowing the VFM teachers to transfer more specialized knowledge related to the downstream data domain. This ensures that the representation biases are further secured in the student model; otherwise the biases could potentially be diminished due to the issue of catastrophic forgetting (French, 1999) during downstream fine-tuning. The overall loss is then formulated as:

$$\mathcal{L}(\boldsymbol{X}) = \gamma \mathcal{L}_{\text{distill}}(\boldsymbol{X}) + \sum_{t \in \mathbb{T}} w_t \mathcal{L}_t(\boldsymbol{X}, \boldsymbol{Y}_t), \tag{6}$$

where $\mathcal{L}_t(\boldsymbol{X}, \boldsymbol{Y}_t)$ is the task-specific loss for task t, computed using the ground truth $\boldsymbol{Y}_t$. The hyperparameter γ balances the distillation loss and the task losses, with a default value of 1.0 for simplicity, while w_t adjusts the importance of each task. We set fixed w_t values following the standard practice in MTL (Maninis et al., 2019; Kanakis et al., 2020).

4 Experiments

4.1 Experimental Setup

Datasets. We conduct experiments on two widely-used multi-task datasets: PASCAL-Context (Mottaghi et al., 2014) with five vision tasks and NYUD-v2 (Silberman et al., 2012) with four tasks. Details can be found in Appendix B.2.

Implementation. We employ a pretrained ViT backbone for TAS and use simple task-specific heads consisting of MLP and convolution layers for decoding. The VFM teachers are DINOv2, CLIP, and SAM with the ViT-L backbones, unless otherwise stated. More implementation details are provided in Appendix B to ensure reproducibility.

Baselines. To evaluate the effectiveness of our method, we consider three categories of baselines: (1) *Single-task baseline*, where individual models are trained for each task using the same ViT-initialized architecture, and *multi-task baseline*, where a shared encoder and task-specific heads are trained jointly. (2) *Multi-teacher VFM distillation approaches*, namely RADIO (Ranzinger et al., 2024b) and Theia (Shang et al., 2024). We use their released models as encoder backbones, coupled with the same task heads as ours. (3) *State-of-the-art MTL models*, which involves complicated encoder or decoder designs. We assess the overall performance of each model with MTL Gain Δ_m by calculating the average relative difference across all tasks compared to the single-task baseline (Maninis et al., 2019).

4.2 Main Results

Figure 3 presents a comparison between our proposed SAK and representative baseline methods on both PASCAL-Context and NYUD-v2 datasets, with all methods using the ViT-B backbones. On PASCAL-Context, SAK greatly boosts the performance in Semseg and Parsing, achieving an overall improvement of 1.66% over the previous SOTA. On NYUD-v2, our method establishes a new milestone across all four tasks, increasing the MTL Gain metric from the previous best of 6.33% to 11.11%. We provide more comprehensive comparisons on both datasets using the ViT-L backbones in Tables 5 and 6, and ViT-S/Swin-S backbones in Appendix C. Our approach consistently outperforms previous methods, achieving the best results on 7 out of 9 tasks and both MTL Gain metrics. Notably, SAK significantly surpasses the SOTAs in MTL (BFCI (Zhang et al., 2023b), MLoRE (Yang et al., 2024d)) by nearly 10% on NYUD-v2, all while using fewer parameters.

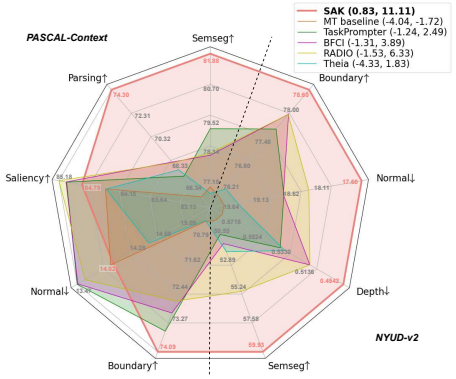

Figure 3: **Performance comparison on two datasets,** based on ViT-B backbones. MTL Gain Δ_m on two datasets are shown in the legend, respectively.

Table 3: **Ablation of proposed modules.** '↑': higher is better; '↓': lower is better; 'Δ_m': MTL Gain w.r.t. single-task baseline. 'Rep Sim' denotes the average cosine similarity between the representations of student and corresponding teachers on the ImageNet-1k validation set.

TSAP	MoR	Rep Sim↑	Semseg mIoU↑	Parsing mIoU↑	Saliency maxF↑	Normal mErr↓	Boundary odsF↑	Δ_m% ↑
✘	✘	0.3344	80.97	69.71	84.64	14.11	72.82	-1.21
✔	✘	0.8708	81.26	69.92	84.31	14.45	71.41	-2.03
✔	✔	0.8708	**81.65**	**72.38**	**84.87**	**14.05**	**73.23**	**-0.03**

Table 4: **Ablation of our two-stage training paradigm.** 'Distill': distillation loss; 'Task': task-specific losses.

Stage1	Stage2		Semseg mIoU↑	Parsing mIoU↑	Saliency maxF↑	Normal mErr↓	Boundary odsF↑	Δ_m% ↑
	Distill	Task						
✘	✘	✔	76.76	65.26	84.39	13.98	70.37	-4.04
✘	✔	✔	77.06	65.08	84.67	**13.83**	70.74	-3.63
✔	✘	✔	80.48	71.16	**85.04**	13.92	72.60	-0.60
✔	✔	✔	**81.65**	**72.38**	84.87	14.05	**73.23**	**-0.03**

Table 5: **Comparison with state of the arts on PASCAL-Context,** based on ViT-L backbones.

Model	Backbone	#Param	Semseg mIoU↑	Parsing mIoU↑	Saliency maxF↑	Normal mErr↓	Boundary odsF↑	Δ_m% ↑
Single-task baseline	ViT-L	1573M	81.61	72.77	83.80	13.87	75.24	0.00
Multi-task baseline	ViT-L	357M	79.26	68.28	84.16	14.06	71.59	-2.97
PAD-Net (Xu et al., 2018)	ViT-L	330M	78.01	67.12	79.21	14.37	72.60	-4.95
MTI-Net (Vandenhende et al., 2020)	ViT-L	851M	78.31	67.40	84.75	14.67	73.00	-3.81
ATRC (Brüggemann et al., 2021)	ViT-L	340M	77.11	66.84	81.20	14.23	72.10	-4.71
InvPT (Ye & Xu, 2022)	ViT-L	423M	79.03	67.61	84.81	14.15	73.00	-2.81
InvPT++ (Ye & Xu, 2024)	ViT-L	421M	80.22	69.12	84.74	13.73	74.20	-1.19
TaskPrompter (Ye & Xu, 2023b)	ViT-L	401M	80.89	68.89	84.83	13.72	73.50	-1.24
TaskExpert (Ye & Xu, 2023a)	ViT-L	420M	80.64	69.42	84.87	13.56	73.30	-0.97
BFCI (Zhang et al., 2023b)	ViT-L	477M	80.64	70.06	84.64	13.82	72.96	-1.32
3D-aware (Li et al., 2024a)	ViT-L	430M	79.53	69.12	84.94	13.53	74.00	-1.08
TSP (Wang et al., 2024b)	ViT-L	423M	81.48	70.64	84.86	13.69	74.80	-0.22
MLoRE (Yang et al., 2024d)	ViT-L	407M	81.41	70.52	84.90	13.51	75.42	0.16
RADIO (Ranzinger et al., 2024b)	ViT-L	372M	81.11	71.50	**85.17**	**13.49**	74.80	0.29
SAK (Ours)	ViT-L	407M	**84.01**	**76.99**	84.65	13.82	**76.27**	**2.30**

4.3 In-depth Analysis

We conduct extensive experiments to validate the effectiveness and generalization of our proposed SAK framework. All experimental analyses are based on the ViT-B backbones for both teachers and student and the PASCAL-Context dataset unless otherwise specified.

Ablation study. An ablation study is conducted to discern the individual contributions of the main components in SAK, namely TSAP and MoR Router, as outlined in Table 3. We consider two model variants: (1) a model without the TSAP and MoR Router modules (row 1), which corresponds to the student distilled naively regardless of representation biases, as studied in Table 2; (2) a model distilled with TSAP in the first stage but trained without MoR Routers in the second stage (row 2), where the biased representations from multiple VFMs are simply added together. Firstly, our results confirm that our proposed TSAP effectively preserves the representation biases from the teachers as indicated by a higher average similarity between the student and teachers. Additionally, we prove that a simple fusion of diverse biased knowledge does not lead to an overall improvement and may even fall behind compared to the student without biases. With the synergization of TSAP and MoR Router, our proposed SAK not only preserves and reproduces the representation biases after distillation but also optimally capitalizes on these biases to maximize multi-task performance. The upper-bound results of teacher amalgamation are presented in Appendix C.

Table 4 reports another ablation on our training paradigm, highlighting the contributions of each stage. The results show that Stage 1, which distills knowledge from VFM teachers on ImageNet, is a primary factor of performance enhancement. Meanwhile, incorporating the distillation loss during Stage 2 consistently boosts final outcomes, regardless of whether Stage 1 is applied. This underscores the effectiveness of transferring specialized knowledge in the downstream data domain.

Table 6: **Comparison with state of the arts on NYUD-v2,** based on ViT-L backbones.

Model	Backbone	#Param	Semseg mIoU↑	Depth RMSE↓	Normal mErr↓	Boundary odsF↑	$\Delta_m\%\uparrow$
Single-task baseline	ViT-L	1259M	54.19	0.5560	19.22	78.09	0.00
Multi-task baseline	ViT-L	346M	52.42	0.5413	19.29	76.50	-0.76
InvPT (Ye & Xu, 2022)	ViT-L	402M	53.56	0.5183	19.04	78.10	1.64
InvPT++ (Ye & Xu, 2024)	ViT-L	∼402M	53.85	0.5096	18.67	78.10	2.65
TaskPrompter (Ye & Xu, 2023b)	ViT-L	392M	55.30	0.5152	18.47	78.20	3.36
TaskExpert (Ye & Xu, 2023a)	ViT-L	400M+	55.35	0.5157	18.54	78.40	3.33
BFCI (Zhang et al., 2023b)	ViT-L	400M+	55.51	0.4930	18.47	78.22	4.46
3D-aware (Li et al., 2024a)	ViT-L	409M	54.87	0.5006	18.55	78.30	3.74
TSP (Wang et al., 2024b)	ViT-L	402M	55.39	0.4961	18.44	77.50	4.07
MLoRE (Yang et al., 2024d)	ViT-L	552M	55.96	0.5076	18.33	78.43	4.26
RADIO (Ranzinger et al., 2024b)	ViT-L	362M	59.32	0.4698	17.46	79.41	8.95
SAK (Ours)	ViT-L	394M	**63.18**	**0.4313**	**16.25**	**79.43**	**14.05**

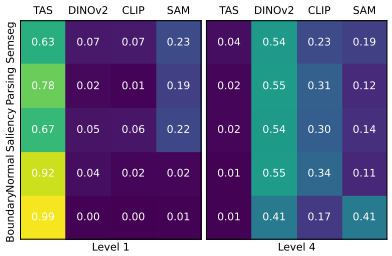

Figure 4: **Weights of different experts learned by MoR Routers.**

Table 7: **Performance w.r.t. different combinations of VFM teachers.** Integrating knowledge from three teachers leads to the strongest overall performance.

Teachers	Semseg mIoU↑	Parsing mIoU↑	Saliency maxF↑	Normal mErr↓	Boundary odsF↑	$\Delta_m\%\uparrow$
Multi-task baseline	76.76	65.26	84.39	13.98	70.37	-4.04
DINOv2	79.05	69.55	84.29	14.07	71.14	-2.21
CLIP	80.12	67.57	83.81	14.41	70.62	-3.26
SAM	63.47	63.99	**85.02**	**13.95**	**73.27**	-6.74
DINOv2+CLIP	81.53	71.95	84.49	14.16	72.59	-0.60
DINOv2+CLIP+SAM	**81.65**	**72.38**	84.87	14.05	73.23	**-0.03**

Impact of teacher selection. To investigate whether knowledge from all teachers can be effectively incorporated into the student model and how each teacher contributes to downstream tasks, we experiment on different combinations of VFM teachers in Table 7. When using a single teacher, SAK effectively learns the teacher's representation bias, as the student distilled from DINOv2 or CLIP performs well in segmentation tasks, while SAM's student is better in tasks requiring finer details. Combining DINOv2 and CLIP continues to improve segmentation tasks, potentially due to their complementary strengths in localized feature learning and semantic understanding. Including SAM further benefits all tasks, leading to the best overall results. We also visualize the gating weights learned by our proposed MoR Routers at the lowest and highest levels in Figure 4. At the lowest level, where VFM teachers share more general knowledge about the details, tasks tend to rely on the shared TAS and SAM's bias. Conversely, the representation biases become more pronounced at higher levels; therefore, the teacher-specific representations are predominantly selected. Further analysis is provided in Appendix D.3 and D.4.

Impact of downstream data size. To assess the robustness of multi-teacher VFM distillation methods, we conduct experiments using varying numbers of samples among {25%, 50%, 75%, 100%} from the downstream dataset. As depicted in Figure 5, while all models show an upward trend as the number of data samples increases, our SAK consistently outperforms the other two distillation baselines across all settings. Particularly, SAK surpasses the second-best method by a clear margin of over 3% in scenarios with substantially fewer samples such as merely 25%.

Scaling with model size. In Table 8, we explore the impact of scaling the backbone sizes of the VFM teachers and student by forming various combinations. The results indicate that increasing the capacity of the student model, while keeping the teacher models fixed (row 1 *vs*. row 2, row 3 *vs*. row 4), yields remarkable improvements across nearly all tasks. Additionally, scaling up the teacher models without altering the student (row 2 *vs*. 3) also proves beneficial. These results demonstrate the versatility and robustness of our approach in adapting to models of varying sizes.

Compatibility with different decoders. It is worth noting that our SAK framework is flexible and does not impose constraints on the design of the backbone, the adapters in TSAP, or the decoder heads. As shown in Table 9, we replace the simple head with the more complex MLoRE decoder (Yang et al., 2024d). Even with a simple head, SAK surpasses MLoRE by 0.8%, and integrating the MLoRE decoder further enhances overall performance by 1.72%.

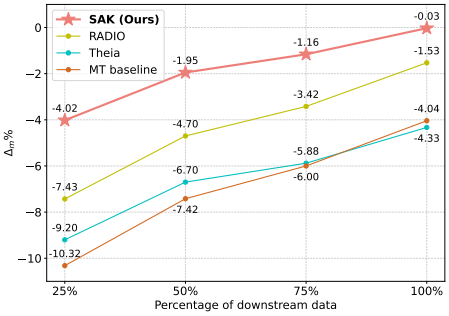

Figure 5: **Performance w.r.t. downstream data percentage.** MTL Gain is computed w.r.t. single-task baseline on full dataset. SAK is the most robust in downstream tasks.

Table 8: **Performance w.r.t. different settings of teacher and student sizes.** SAK shows robustness across teachers or students with varying capacities.

Backbone		Semseg	Parsing	Saliency	Normal	Boundary
Teachers	Student	mIoU↑	mIoU↑	maxF↑	mErr↓	odsF↑
ViT-B	ViT-S	78.66	68.46	84.66	14.33	70.28
ViT-B	ViT-B	81.65	72.38	84.87	14.05	73.23
ViT-L	ViT-B	81.88	74.30	84.79	14.02	74.09
ViT-L	ViT-L	84.01	76.99	84.65	13.82	76.27

Table 9: **Performance of SAK integrated with MLoRE.** SAK further benefits from stronger decoders.

Enc.	Dec.	Semseg mIoU↑	Parsing mIoU↑	Saliency maxF↑	Normal mErr↓	Boundary odsF↑	$\Delta_m\%\uparrow$
ViT	MLoRE	79.26	67.82	**85.31**	**13.65**	74.69	-0.83
SAK	Simple	81.65	72.38	84.87	14.05	73.23	-0.03
SAK	MLoRE	**82.74**	**74.28**	84.58	13.89	**75.96**	**1.69**

5 Related Work

Knowledge Distillation of Vision Foundation Models. As large-scale generalists, Vision Foundation Models (VFMs) show superior performance in various tasks with minimal tuning, such as CLIP (Radford et al., 2021) for vision-language tasks, DINOv2 (Oquab et al., 2024) for fine-grained recognition, and SAM (Kirillov et al., 2023) for promptable segmentation. To reduce their computational demands while preserving performance, knowledge distillation (Buciluǎ et al., 2006; Hinton et al., 2014) has been widely adopted in compressing VFMs (Vemulapalli et al., 2024; Sun et al., 2023; Yang et al., 2024a). More recently, multiple VFMs are distilled into a single student to combine their strengths: SAM-CLIP (Wang et al., 2024a) merges CLIP into SAM via continual learning and distillation. RADIO (Ranzinger et al., 2024b) learns from CLIP, DINOv2, and SAM to enhance performance on downstream tasks. Theia (Shang et al., 2024) further incorporates Depth Anything (Yang et al., 2024b), showing advantages in robot learning. Different from the straightforward distillation in these methods, we adaptively transfer knowledge from multiple teachers while retaining the unique representation biases to maximize their strengths for multiple tasks.

Multi-Task Learning. Multi-Task Learning (MTL) aims to train a single model capable of handling multiple tasks simultaneously (Caruana, 1997; Zhang & Yang, 2021; Yu et al., 2024). MTL research primarily falls into two categories: multi-task optimization (Kendall et al., 2018; Chen et al., 2018; Yu et al., 2020) and model architecture design (Long et al., 2017; Wallingford et al., 2022; Lu et al., 2024c). Considering vision tasks, most works center on designing architectures, which is further divided into encoder-focused and decoder-focused methods (Vandenhende et al., 2021). Encoder-focused methods develop encoders to extract features for different tasks (Misra et al., 2016; Ruder et al., 2019; Gao et al., 2019), while decoder-focused methods introduce task-interaction modules in decoder to better capture task-specific features (Ye & Xu, 2022; Xu et al., 2023c; Ye & Xu, 2023b).

Knowledge distillation has also been applied to enhance MTL (Li & Bilen, 2020; Jacob et al., 2023; Ghiasi et al., 2021; Luo et al., 2020; Ye et al., 2019a). These methods train a multi-task model to mimic multiple single-task teachers, allowing the student to gain richer information. Xu et al. (2023d) propose directly distilling a small multi-task student from a large multi-task teacher. To the best of our knowledge, our work is the first exploration of multi-task distillation with general-purpose knowledge from task-agnostic VFM teachers, as opposed to task-related teachers trained on target datasets.

6 Conclusion

Building on our analysis of the representation biases in VFMs, we introduce a novel framework SAK, designed to improve multi-task learning by exploiting the complementary biases of multiple VFMs. Through the integration of a Teacher-Agnostic Stem, Teacher-Specific Adapter Paths, and Mixture-of-Representations Routers, SAK effectively preserves the unique representation biases during distillation, thereby enhancing both accuracy and robustness across multiple downstream tasks. Our work opens possibilities for including more advanced teachers and students, and provides a solid foundation for future advancements in multi-task visual learning with foundation models.

ACKNOWLEDGMENTS

This work was supported in part by NSF Grant 2106825, NIFA Award 2020-67021-32799, the Toyota Research Institute, the IBM-Illinois Discovery Accelerator Institute, the Amazon-Illinois Center on AI for Interactive Conversational Experiences, Snap Inc., and the Jump ARCHES endowment through the Health Care Engineering Systems Center at Illinois and the OSF Foundation. This work used computational resources, including the NCSA Delta and DeltaAI supercomputers through allocations CIS230012 and CIS240428 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, as well as the TACC Frontera supercomputer and Amazon Web Services (AWS) through the National Artificial Intelligence Research Resource (NAIRR) Pilot.

ETHICS STATEMENT

Our research adheres to high ethical standards in machine learning and computer vision, ensuring transparency, reproducibility, and fairness in all experiments. While our approach shows promising results, it shares common limitations with other foundation models, particularly regarding data usage. We utilize publicly available datasets in compliance with relevant legal and ethical guidelines, emphasizing proper data handling during selection and pre-processing.

REPRODUCIBILITY STATEMENT

We ensure reproducibility by elaborating implementation details, which includes details on models, datasets, and training setups in Appendix B. Our code and models are publicly available at https://github.com/innovator-zero/SAK.

REFERENCES

Ahmed Agiza, Marina Neseem, and Sherief Reda. MTLoRA: Low-rank adaptation approach for efficient multi-task learning. In *CVPR*, 2024.

Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D. Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *CVPR*, 2019.

Abhishek Aich, Samuel Schulter, Amit K. Roy-Chowdhury, Manmohan Chandraker, and Yumin Suh. Efficient controllable multi-task architectures. In *ICCV*, 2023.

Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In *NeurIPS*, 2014.

Zhipeng Bao, Martial Hebert, and Yu-Xiong Wang. Generative modeling for multi-task visual learning. In *ICML*, 2022.

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

Felix J.S. Bragman, Ryutaro Tanno, Sebastien Ourselin, Daniel C. Alexander, and Jorge Cardoso. Stochastic filter groups for multi-task CNNs: Learning specialist and generalist convolution kernels. In *ICCV*, 2019.

David Brüggemann, Menelaos Kanakis, Anton Obukhov, Stamatios Georgoulis, and Luc Van Gool. Exploring relational context for multi-task dense prediction. In *ICCV*, 2021.

David Brüggemann, Menelaos Kanakis, Stamatios Georgoulis, and Luc Van Gool. Automated search for resource-efficient branched multi-task networks. In *BMVC*, 2020.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *SIGKDD*, 2006.

Shengcao Cao, Mengtian Li, James Hays, Deva Ramanan, Yu-Xiong Wang, and Liang-Yan Gui. Learning lightweight object detectors via multi-teacher progressive distillation. In *ICML*, 2023.

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

Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. In *NeurIPS*, 2022.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *ICML*, 2018.

Zhao Chen, Jiquan Ngiam, Yanping Huang, Thang Luong, Henrik Kretzschmar, Yuning Chai, and Dragomir Anguelov. Just pick a sign: Optimizing deep multitask models with gradient sign dropout. In *NeurIPS*, 2020.

Zitian Chen, Yikang Shen, Mingyu Ding, Zhenfang Chen, Hengshuang Zhao, Erik G. Learned-Miller, and Chuang Gan. Mod-Squad: Designing mixtures of experts as modular multi-task learners. In *CVPR*, 2023.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, 2023.

Michael Crawshaw. Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*, 2020.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *ICLR*, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander T. Toshev, and Vaishaal Shankar. Data filtering networks. In *ICLR*, 2024.

Robert M. French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999.

Takashi Fukuda, Masayuki Suzuki, Gakuto Kurata, Samuel Thomas, Jia Cui, and Bhuvana Ramabhadran. Efficient knowledge distillation from an ensemble of teachers. In *Interspeech*, 2017.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei W. Koh, Olga Saukh, Alexander J. Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. DataComp: In search of the next generation of multimodal datasets. In *NeurIPS*, 2023.

Yuan Gao, Jiayi Ma, Mingbo Zhao, Wei Liu, and Alan Yuille. NDDR-CNN: Layerwise feature fusing in multi-task CNNs by neural discriminative dimensionality reduction. In *CVPR*, 2019.

Golnaz Ghiasi, Barret Zoph, Ekin D. Cubuk, Quoc V. Le, and Tsung-Yi Lin. Multi-task self-training for learning general representations. In *ICCV*, 2021.

Michelle Guo, Albert Haque, De-An Huang, Serena Yeung, and Li Fei-Fei. Dynamic task prioritization for multitask learning. In *ECCV*, 2018.

Pengsheng Guo, Chen-Yu Lee, and Daniel Ulbricht. Learning to branch for multi-task learning. In *ICML*, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

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

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NeurIPS Deep Learning Workshop*, 2014.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *ICML*, 2019.

Huimin Huang, Yawen Huang, Lanfen Lin, Ruofeng Tong, Yen-Wei Chen, Hao Zheng, Yuexiang Li, and Yefeng Zheng. Going beyond multi-task dense prediction with synergy embedding models. In *CVPR*, 2024a.

Siyu Huang, Xi Li, Zhi-Qi Cheng, Zhongfei Zhang, and Alexander Hauptmann. GNAS: A greedy neural architecture search method for multi-attribute learning. In *ACM MM*, 2018.

Suizhi Huang, Shalayiding Sirejiding, Yuxiang Lu, Yue Ding, Leheng Liu, Hui Zhou, and Hongtao Lu. YOLO-Med: Multi-task interaction network for biomedical images. In *ICASSP*, 2024b.

Keishi Ishihara, Anssi Kanervisto, Jun Miura, and Ville Hautamaki. Multi-task learning with attention for end-to-end autonomous driving. In *CVPR*, 2021.

Geethu Miriam Jacob, Vishal Agarwal, and Björn Stenger. Online knowledge distillation for multi-task learning. In *WACV*, 2023.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Adrián Javaloy and Isabel Valera. RotoGrad: Gradient homogenization in multitask learning. In *ICLR*, 2022.

Menelaos Kanakis, David Bruggemann, Suman Saha, Stamatios Georgoulis, Anton Obukhov, and Luc Van Gool. Reparameterizing convolutions for incremental multi-task learning without task interference. In *ECCV*, 2020.

Oğuzhan Fatih Kar, Alessio Tonioni, Petra Poklukar, Achin Kulshrestha, Amir Zamir, and Federico Tombari. BRAVE: Broadening the visual encoding of vision-language models. In *ECCV*, 2024.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, 2018.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *ICCV*, 2023.

Wei-Hong Li and Hakan Bilen. Knowledge distillation for multi-task learning. In *ECCV Workshops*, 2020.

Wei-Hong Li, Steven McDonagh, Ales Leonardis, and Hakan Bilen. Multi-task learning with 3D-aware regularization. In *ICLR*, 2024a.

Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation. *IEEE Transactions on Image Processing*, 2024b.

Hanxue Liang, Zhiwen Fan, Rishov Sarkar, Ziyu Jiang, Tianlong Chen, Kai Zou, Yu Cheng, Cong Hao, and Zhangyang Wang. M^3ViT: Mixture-of-experts vision transformer for efficient multi-task learning with model-accelerator co-design. In *NeurIPS*, 2022.

Baijiong Lin, Weisen Jiang, Pengguang Chen, Yu Zhang, Shu Liu, and Ying-Cong Chen. MT-Mamba: Enhancing multi-task dense scene understanding by mamba-based decoders. In *ECCV*, 2024.

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, Jiaming Han, Siyuan Huang, Yichi Zhang, Xuming He, Hongsheng Li, and Yu Qiao. SPHINX: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *NeurIPS*, 2021a.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2024b.

Liyang Liu, Yi Li, Zhanghui Kuang, Jing-Hao Xue, Yimin Chen, Wenming Yang, Qingmin Liao, and Wayne Zhang. Towards impartial multi-task learning. In *ICLR*, 2021b.

Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *CVPR*, 2019.

Yen-Cheng Liu, Chih-Yao Ma, Junjiao Tian, Zijian He, and Zsolt Kira. Polyhistor: Parameter-efficient multi-task adaptation for dense vision tasks. In *NeurIPS*, 2022.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021c.

Zichang Liu, Qingyun Liu, Yuening Li, Liang Liu, Anshumali Shrivastava, Shuchao Bi, Lichan Hong, Ed H. Chi, and Zhe Zhao. Wisdom of committee: Distilling from foundation model to specialized application model. *arXiv preprint arXiv:2402.14035*, 2024c.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Philip S. Yu. Learning multiple tasks with multilinear relationship networks. In *NeurIPS*, 2017.

Yongxi Lu, Abhishek Kumar, Shuangfei Zhai, Yu Cheng, Tara Javidi, and Rogerio Feris. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In *CVPR*, 2017.

Yuxiang Lu, Suizhi Huang, Yuwen Yang, Shalayiding Sirejiding, Yue Ding, and Hongtao Lu. FedHCA2: Towards hetero-client federated multi-task learning. In *CVPR*, 2024a.

Yuxiang Lu, Shalayiding Sirejiding, Bayram Bayramli, Suizhi Huang, Yue Ding, and Hongtao Lu. Task indicating transformer for task-conditional dense predictions. In *ICASSP*, 2024b.

Yuxiang Lu, Shalayiding Sirejiding, Yue Ding, Chunlin Wang, and Hongtao Lu. Prompt guided transformer for multi-task dense prediction. *IEEE Transactions on Multimedia*, 2024c.

Sihui Luo, Xinchao Wang, Gongfan Fang, Yao Hu, Dapeng Tao, and Mingli Song. Knowledge amalgamation from heterogeneous networks by common feature learning. In *IJCAI*, 2019.

Sihui Luo, Wenwen Pan, Xinchao Wang, Dazhou Wang, Haihong Tang, and Mingli Song. Collaboration by competition: Self-coordinated knowledge amalgamation for multi-talent student learning. In *ECCV*, 2020.

Arun Mallya, Dillon Davis, and Svetlana Lazebnik. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In *ECCV*, 2018.

Yunze Man, Shuhong Zheng, Zhipeng Bao, Martial Hebert, Liang-Yan Gui, and Yu-Xiong Wang. Lexicon3D: Probing visual foundation models for complex 3D scene understanding. In *NeurIPS*, 2024.

Kevis-Kokitsi Maninis, Ilija Radosavovic, and Iasonas Kokkinos. Attentive single-tasking of multiple tasks. In *CVPR*, 2019.

David R. Martin, Charless C. Fowlkes, and Jitendra Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *TPAMI*, 26(5):530–549, 2004.

Otniel-Bogdan Mercea, Alexey Gritsenko, Cordelia Schmid, and Anurag Arnab. Time- Memory- and Parameter-Efficient Visual Adaptation. In *CVPR*, 2024.

Elliot Meyerson and Risto Miikkulainen. Beyond shared hierarchies: Deep multitask learning through soft layer ordering. In *ICLR*, 2018.

Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *CVPR*, 2016.

Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, 2014.

Marina Neseem, Ahmed Agiza, and Sherief Reda. AdaMTL: Adaptive input-dependent inference for efficient multi-task learning. In *CVPR*, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024.

Ziqi Pang, Ziyang Xie, Yunze Man, and Yu-Xiong Wang. Frozen transformers in language models are effective visual encoder layers. In *ICLR*, 2024.

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

Jordi Pont-Tuset and Ferran Marques. Supervised evaluation of image segmentation and object proposal techniques. *TPAMI*, 38(7):1465–1478, 2015.

Han Qiu, Jiaxing Huang, Peng Gao, Lewei Lu, Xiaoqin Zhang, and Shijian Lu. Masked AutoDecoder is effective multi-task vision generalist. In *CVPR*, 2024.

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. In *ICML*, 2021.

Mike Ranzinger, Jon Barker, Greg Heinrich, Pavlo Molchanov, Bryan Catanzaro, and Andrew Tao. PHI-S: Distribution balancing for label-free multi-teacher distillation. *arXiv preprint arXiv:2410.01680*, 2024a.

Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. AM-RADIO: Agglomerative vision foundation model reduce all domains into one. In *CVPR*, 2024b.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *ICLR*, 2025.

Dripta S. Raychaudhuri, Yumin Suh, Samuel Schulter, Xiang Yu, Masoud Faraki, Amit K. Roy-Chowdhury, and Manmohan Chandraker. Controllable dynamic multi-task architectures. In *CVPR*, 2022.

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

Clemens Rosenbaum, Tim Klinger, and Matthew Riemer. Routing networks: Adaptive selection of non-linear functions for multi-task learning. In *ICLR*, 2018.

Karsten Roth, Lukas Thede, A. Sophia Koepke, Oriol Vinyals, Olivier J. Henaff, and Zeynep Akata. Fantastic gains and where to find them: On the existence and prospect of general knowledge transfer between any pretrained model. In *ICLR*, 2024.

Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, and Anders Søgaard. Latent multi-task architecture learning. In *AAAI*, 2019.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Mert Bulent Sariyildiz, Philippe Weinzaepfel, Thomas Lucas, Diane Larlus, and Yannis Kalantidis. UNIC: Universal classification models via multi-teacher distillation. In *ECCV*, 2024.

Bharat Bhusan Sau and Vineeth N Balasubramanian. Deep model compression: Distilling knowledge from noisy teachers. *arXiv preprint arXiv:1610.09650*, 2016.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *NeurIPS*, 2018.

Jinghuan Shang, Karl Schmeckpeper, Brandon B. May, Maria Vittoria Minniti, Tarik Kelestemur, David Watkins, and Laura Herlant. Theia: Distilling diverse vision foundation models for robot learning. In *CoRL*, 2024.

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*, 2017.

Chengchao Shen, Xinchao Wang, Jie Song, Li Sun, and Mingli Song. Amalgamating knowledge towards comprehensive classification. In *AAAI*, 2019a.

Chengchao Shen, Mengqi Xue, Xinchao Wang, Jie Song, Li Sun, and Mingli Song. Customizing student networks from heterogeneous teachers via adaptive knowledge amalgamation. In *ICCV*, 2019b.

Michal Shlapentokh-Rothman, Ansel Blume, Yao Xiao, Yuqun Wu, Sethuraman TV, Heyi Tao, Jae Yong Lee, Wilfredo Torres, Yu-Xiong Wang, and Derek Hoiem. Region-Based representations revisited. In *CVPR*, 2024.

Sara Shoouri, Mingyu Yang, Zichen Fan, and Hun-Seok Kim. Efficient computation sharing for multi-task visual scene understanding. In *ICCV*, 2023.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV*, 2012.

Shalayiding Sirejiding, Yuxiang Lu, Hongtao Lu, and Yue Ding. Scale-aware task message transferring for multi-task learning. In *ICME*, 2023.

Shalayiding Sirejiding, Bayram Bayramli, Yuxiang Lu, Suizhi Huang, Hongtao Lu, and Yue Ding. Adaptive task-wise message passing for multi-task learning: A spatial interaction perspective. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024a.

Shalayiding Sirejiding, Bayram Bayramli, Yuxiang Lu, Yuwen Yang, Tamam Alsarhan, Hongtao Lu, and Yue Ding. Task-Interaction-Free multi-task learning with efficient hierarchical feature representation. In *ACM MM*, 2024b.

Yanfei Songa, Bangzheng Pua, Peng Wanga, Hongxu Jiang, Dong Donga, and Yiqing Shen. SAM-Lightening: A lightweight segment anything model with dilated flash attention to achieve 30 times acceleration. *arXiv preprint arXiv:2403.09195*, 2024.

Guolei Sun, Thomas Probst, Danda Pani Paudel, Nikola Popović, Menelaos Kanakis, Jagruti Patel, Dengxin Dai, and Luc Van Gool. Task switching network for multi-task learning. In *ICCV*, 2021.

Ximeng Sun, Pengchuan Zhang, Peizhao Zhang, Hardik Shah, Kate Saenko, and Xide Xia. DIME-FM: Distilling multimodal and efficient foundation models. In *ICCV*, 2023.

Jidapa Thadajarassiri, Thomas Hartvigsen, Xiangnan Kong, and Elke A Rundensteiner. Semi-supervised knowledge amalgamation for sequence classification. In *AAAI*, 2021.

Jidapa Thadajarassiri, Thomas Hartvigsen, Walter Gerych, Xiangnan Kong, and Elke Rundensteiner. Knowledge amalgamation for multi-label classification via label dependency transfer. In *AAAI*, 2023.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *NeurIPS*, 2024a.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs. In *CVPR*, 2024b.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021.

Simon Vandenhende, Stamatios Georgoulis, Bert De Brabandere, and Luc Van Gool. Branched multi-task networks: Deciding what layers to share. In *BMVC*, 2019.

Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. MTI-Net: Multi-scale task interaction networks for multi-task learning. In *ECCV*, 2020.

Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Multi-task learning for dense prediction tasks: A survey. *TPAMI*, 44(7): 3614–3633, 2021.

Raviteja Vemulapalli, Hadi Pouransari, Fartash Faghri, Sachin Mehta, Mehrdad Farajtabar, Mohammad Rastegari, and Oncel Tuzel. Knowledge transfer from vision foundation models for efficient training of small task-specific models. In *ICML*, 2024.

Jayakorn Vongkulbhisal, Phongtharin Vinayavekhin, and Marco Visentini-Scarzanella. Unifying heterogeneous classifiers with distillation. In *CVPR*, 2019.

Matthew Wallingford, Hao Li, Alessandro Achille, Avinash Ravichandran, Charless Fowlkes, Rahul Bhotika, and Stefano Soatto. Task adaptive parameter sharing for multi-task learning. In *CVPR*, 2022.

Ao Wang, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. RepViT-SAM: Towards real-time segmenting anything. *arXiv preprint arXiv:2312.05760*, 2023.

Haoxiang Wang, Pavan Kumar Anasosalu Vasu, Fartash Faghri, Raviteja Vemulapalli, Mehrdad Farajtabar, Sachin Mehta, Mohammad Rastegari, Oncel Tuzel, and Hadi Pouransari. SAM-CLIP: Merging vision foundation models towards semantic and spatial understanding. In *CVPR*, 2024a.

Shuo Wang, Jing Li, Zibo Zhao, Dongze Lian, Binbin Huang, Xiaomei Wang, Zhengxin Li, and Shenghua Gao. TSP-Transformer: Task-specific prompts boosted transformer for holistic scene understanding. In *WACV*, 2024b.

Xuehao Wang, Feiyang Ye, and Yu Zhang. Task-aware low-rank adaptation of segment anything model. *arXiv preprint arXiv:2403.10971*, 2024c.

Zirui Wang, Yulia Tsvetkov, Orhan Firat, and Yuan Cao. Gradient vaccine: Investigating and improving multi-task optimization in massively multilingual models. In *ICLR*, 2021.

Haitao Wen, Lili Pan, Yu Dai, Heqian Qiu, Lanxiao Wang, Qingbo Wu, and Hongliang Li. Class incremental learning with multi-teacher distillation. In *CVPR*, 2024.

Ross Wightman. PyTorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Yi Xin, Junlong Du, Qiang Wang, Zhiwen Lin, and Ke Yan. VMT-Adapter: Parameter-efficient transfer learning for multi-task dense scene understanding. In *AAAI*, 2024a.

Zewei Xin, Shalayiding Sirejiding, Yuxiang Lu, Yue Ding, Chunlin Wang, Tamam Alsarhan, and Hongtao Lu. TFUT: Task fusion upward transformer model for multi-task learning on dense prediction. *Computer Vision and Image Understanding*, 244:104014, 2024b.

Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. PAD-Net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *CVPR*, 2018.

Xiaogang Xu, Hengshuang Zhao, Vibhav Vineet, Ser-Nam Lim, and Antonio Torralba. MTFormer: Multi-task learning via transformer and cross-task reasoning. In *ECCV*, 2022.

Yangyang Xu, Xiangtai Li, Haobo Yuan, Yibo Yang, and Lefei Zhang. Multi-task learning with multi-query transformer for dense prediction. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023a.

Yangyang Xu, Yibo Yang, Bernard Ghanem, Lefei Zhang, Du Bo, and Dacheng Tao. Deformable mixer transformer with gating for multi-task learning of dense prediction. *arXiv preprint arXiv:2308.05721*, 2023b.

Yangyang Xu, Yibo Yang, and Lefei Zhang. DeMT: Deformable mixer transformer for multi-task learning of dense prediction. In *AAAI*, 2023c.

Yangyang Xu, Yibo Yang, and Lefei Zhang. Multi-task learning with knowledge distillation for dense prediction. In *ICCV*, 2023d.

Chuanguang Yang, Zhulin An, Libo Huang, Junyu Bi, Xinqiang Yu, Han Yang, Boyu Diao, and Yongjun Xu. CLIP-KD: An empirical study of clip model distillation. In *CVPR*, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024b.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything V2. In *NeurIPS*, 2024c.

Siwei Yang, Hanrong Ye, and Dan Xu. Contrastive multi-task dense prediction. In *AAAI*, 2023.

Yongxin Yang and Timothy Hospedales. Deep multi-task representation learning: A tensor factorisation approach. In *ICLR*, 2017.

Yuqi Yang, Peng-Tao Jiang, Qibin Hou, Hao Zhang, Jinwei Chen, and Bo Li. Multi-task dense prediction via mixture of low-rank experts. In *CVPR*, 2024d.

Feiyang Ye, Baijiong Lin, Zhixiong Yue, Pengxin Guo, Qiao Xiao, and Yu Zhang. Multi-objective meta learning. In *NeurIPS*, 2021.

Hanrong Ye and Dan Xu. Inverted pyramid multi-task transformer for dense scene understanding. In *ECCV*, 2022.

Hanrong Ye and Dan Xu. TaskExpert: Dynamically assembling multi-task representations with memorial mixture-of-experts. In *ICCV*, 2023a.

Hanrong Ye and Dan Xu. TaskPrompter: Spatial-channel multi-task prompting for dense scene understanding. In *ICLR*, 2023b.

Hanrong Ye and Dan Xu. InvPT++: Inverted pyramid multi-task transformer for visual scene understanding. *TPAMI*, 2024.

Jingwen Ye, Yixin Ji, Xinchao Wang, Kairi Ou, Dapeng Tao, and Mingli Song. Student becoming the master: Knowledge amalgamation for joint scene parsing, depth estimation, and more. In *CVPR*, 2019a.

Jingwen Ye, Xinchao Wang, Yixin Ji, Kairi Ou, and Mingli Song. Amalgamating filtered knowledge: learning task-customized student from multi-task teachers. In *IJCAI*, 2019b.

Shan You, Chang Xu, Chao Xu, and Dacheng Tao. Learning from multiple teacher networks. In *SIGKDD*, 2017.

Jun Yu, Yutong Dai, Xiaokang Liu, Jin Huang, Yishan Shen, Ke Zhang, Rong Zhou, Eashan Adhikarla, Wenxuan Ye, Yixin Liu, Zhaoming Kong, Kai Zhang, Yilong Yin, Vinod Namboodiri, Brian D. Davison, Jason H. Moore, and Yong Chen. Unleashing the power of multi-task learning: A comprehensive survey spanning traditional, deep, and pretrained foundation model eras. *arXiv preprint arXiv:2404.18961*, 2024.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *NeurIPS*, 2020.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023.

Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023a.

Jingdong Zhang, Jiayuan Fan, Peng Ye, Bo Zhang, Hancheng Ye, Baopu Li, Yancheng Cai, and Tao Chen. Rethinking of feature interaction for multi-task learning on dense prediction. *arXiv preprint arXiv:2312.13514*, 2023b.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023c.

Xiaoya Zhang, Ling Zhou, Yong Li, Zhen Cui, Jin Xie, and Jian Yang. Transfer vision patterns for multi-task pixel learning. In *ACM MM*, 2021.

Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, 2021.

Zhenyu Zhang, Zhen Cui, Chunyan Xu, Zequn Jie, Xiang Li, and Jian Yang. Joint task-recursive learning for semantic segmentation and depth estimation. In *ECCV*, 2018.

Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In *CVPR*, 2019.

Zhihao Zhang, Shengcao Cao, and Yu-Xiong Wang. TAMM: TriAdapter multi-modal learning for 3D shape understanding. In *CVPR*, 2024a.

Zhuoyang Zhang, Han Cai, and Song Han. EfficientViT-SAM: Accelerated segment anything model without performance loss. *arXiv preprint arXiv:2402.05008*, 2024b.

Chong Zhou, Xiangtai Li, Chen Change Loy, and Bo Dai. EdgeSAM: Prompt-in-the-loop distillation for on-device deployment of SAM. *arXiv preprint arXiv:2312.06660*, 2023.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image BERT pre-training with online tokenizer. In *ICLR*, 2022.

Ling Zhou, Zhen Cui, Chunyan Xu, Zhenyu Zhang, Chaoqun Wang, Tong Zhang, and Jian Yang. Pattern-structure diffusion for multi-task learning. In *CVPR*, 2020.

Zhuofan Zong, Bingqi Ma, Dazhong Shen, Guanglu Song, Hao Shao, Dongzhi Jiang, Hongsheng Li, and Yu Liu. MoVA: Adapting mixture of vision experts to multimodal context. In *NeurIPS*, 2024.