VL2Lite: Task-Specific Knowledge Distillation from Large Vision-Language Models to Lightweight Networks

Jinseong Jang Chunfei Ma Byeongwon Lee
Vision Lab, AI R&D Center, SK Telecom
{jsjang0814, chunfei.ma, bwon.lee}@sk.com

Abstract

Deploying high-performing neural networks in resource-constrained environments poses a significant challenge due to the computational demands of large-scale models. We introduce VL2Lite, a knowledge distillation framework designed to enhance the performance of lightweight neural networks in image classification tasks by leveraging the rich representational knowledge from Vision-Language Models (VLMs). VL2Lite directly integrates multi-modal knowledge from VLMs into compact models during training, effectively compensating for the limited computational and modeling capabilities of smaller networks. By transferring high-level features and complex data representations, our approach improves the accuracy and efficiency of image classification tasks without increasing computational overhead during inference. Experimental evaluations demonstrate that VL2Lite achieves up to a 7% improvement in classification performance across various datasets. This method addresses the challenge of deploying accurate models in environments with constrained computational resources, offering a balanced solution between model complexity and operational efficiency.

1. Introduction

Large-scale Vision-Language Models (VLMs) have significantly advanced the field by effectively integrating high-dimensional visual data with rich textual information [1, 2, 23, 24, 37, 44, 45, 48, 52]. Trained on extensive multimodal datasets [8, 40], these models excel in various downstream tasks, from image recognition [32] to text-to-image generation [38]. However, their practical deployment is hindered by substantial computational complexity [50], posing challenges in real-time applications and resource-constrained environments like edge devices.

Lightweight neural networks aim to alleviate this issue by reducing computational demands but often suffer from performance degradation compared to their large-scale counterparts [9, 50]. Enhancing the efficiency of lightweight models without sacrificing accuracy remains a critical challenge in computer vision.

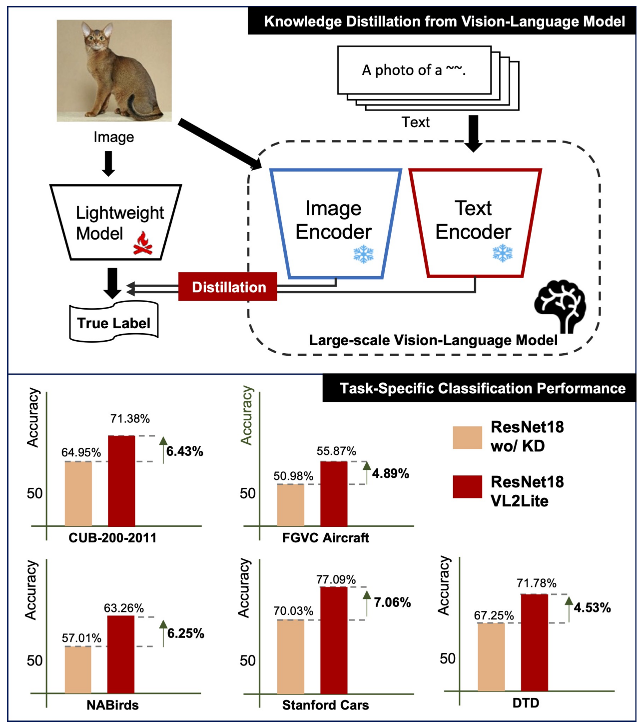

Figure 1. An overview of our proposed method **VL2Lite**, illustrating the knowledge distillation from a VLM to a lightweight neural network. The bar charts on the right demonstrate the task-specific classification performance improvement.

Knowledge Distillation (KD) [10, 15, 17, 28] has emerged as an effective technique to transfer knowledge from a large teacher model to a smaller student model, improving performance while maintaining computational efficiency. Traditional KD methods involve a two-stage training process—pre-training the teacher model on specific tasks, followed by distillation to the student model—which is time-consuming and resource-intensive. Moreover, these methods risk transferring the teacher's biases to the student [34].

In this paper, we propose **VL2Lite**, a novel onestep knowledge distillation framework that directly transfers multimodal knowledge from pre-trained VLMs to lightweight neural networks during task-specific training. By leveraging the rich visual and linguistic representations inherent in VLMs without additional teacher training, our approach streamlines the training pipeline and enhances the student model's representational capacity. This is analogous to fine-tuning small models using pre-trained Large Language Models (LLMs) [18, 20, 46], simplifying the overall process.

Furthermore, we employ prompt engineering to extract meaningful text features, augmenting the student model's linguistic understanding and improving performance on visual tasks. The proposed method integrates task loss, visual KD loss, and linguistic KD loss into a composite loss function, enabling simultaneous classification and knowledge distillation in a single training phase.

To bridge the gap between the high-dimensional feature space of VLMs and the lower-dimensional space of lightweight models, we introduce a knowledge condensation layer. This layer effectively compresses the rich representations from VLMs into a format suitable for the student model, facilitating efficient knowledge transfer.

Our method outperforms existing supervised and KD techniques on multiple classification tasks, with gains up to 7%. This demonstrates the effectiveness of VL2Lite for real-time deployment on resource-constrained devices in Fig. 1. Our main contributions are:

1) We introduce a novel approach for direct KD from VLMs to lightweight models to reduce time and computational resources.

2) Our proposed VL2Lite demonstrates an enhanced ability to distill extensive multi-modal knowledge more effectively than existing KD techniques, ensuring that lightweight models retain a high level of performance while operating under computational constraints.

3) By employing text representation features obtained via prompt engineering, we enhance the linguistic comprehension of lightweight models, which leads to a significant boost in their visual task performance.

4) The proposed distillation approach surpasses traditional supervised learning and KD methodologies in various task-specific classifications, showing up to a 7% higher performance across multiple datasets. This demonstrates its robustness and versatility in enhancing model performance for real-time processing and on-edge computing devices with limited resources.

2. Related Work

2.1. Vision-Language Models

Large-scale Vision-Language Models (VLMs) such as CLIP [37], OpenCLIP [2], GLIP [24], OFA [44], SimVLM [48], BEiT [45], Florence [52], Flamingo [1], and LLaVa [27] have significantly advanced multimodal learning by effectively integrating visual and textual modalities. Leveraging massive internet-scale datasets [8, 40], these models demonstrate exceptional performance across various tasks such as image classification, object detection, and image-to-text generation.

Despite their impressive capabilities, VLMs often encounter challenges in fine-grained, task-specific distinctions within categories, particularly when such nuances are subtle or underrepresented in training data. To enhance task-specific performance, various methods have been proposed [13, 26, 51, 53]. Techniques like prompt engineering [12, 13] involve designing text prompts to guide VLM outputs toward desired responses. Moreover, fine-tuning approaches that integrate vision-language strategies have proven effective in optimizing VLMs for specialized applications.

In contrast to these methods, our study introduces a distillation framework that transfers the rich feature representations of large-scale VLMs into lightweight neural network architectures. We focus on effectively distilling both visual and linguistic representations into compact models, enabling deployment in resource-constrained environments while preserving the adaptability and robustness inherent in extensively trained multimodal models.

2.2. Knowledge Distillation to Lightweight Models

Knowledge Distillation (KD) [16, 47] is a technique for transferring knowledge from a larger, well-trained teacher model to a smaller, more efficient student model. Traditional KD approaches align the student's outputs with those of the teacher, often using Kullback-Leibler divergence to match predicted probabilities. Recent advancements extend beyond output alignment to transferring intermediate feature representations and relational knowledge [6, 9, 35, 47], enhancing the student's ability to replicate the teacher's data processing and interpretation mechanisms.

Conventional KD typically involves a two-stage process: pre-training a large teacher model and then distilling knowledge to the student model. This approach is time-consuming and resource-intensive, and may propagate biases from the teacher to the student [34]. Our proposed method deviates from this by employing a single-stage process where only the student model is trained, using a fixed pre-trained teacher model. This strategy prevents bias propagation and allows the teacher's guidance to regularize the student model, making the distillation process more resilient to biases in the training data.

Unlike two-stage methods that often fine-tune the teacher on the same target dataset before distilling, our approach keeps the large VLM frozen, bypassing teacher fine-tuning. Thus, we only train the student (plus minimal MLP layers) in a single step, reducing overhead and risk of overfitting the teacher.

This efficiency is crucial for deployment in resource-constrained environments, setting our approach apart from traditional methods. By freezing the teacher model, VL2Lite minimizes computational overhead and reduces the transfer of dataset-specific biases, ensuring a more balanced learning process.

2.3. Knowledge Distillation from Vision-Language Models

With the advent of VLMs, knowledge distillation has expanded to encompass multimodal knowledge transfer, integrating both visual and textual data [5, 49]. This extension allows for more robust knowledge transfer to student models, enhancing performance on tasks such as open-vocabulary detection. The integration of self-supervised learning with KD has introduced methods for transferring relational knowledge across modalities [6, 9], using techniques like contrastive loss to bolster the student's learning capacity.

Recent efforts have focused on condensing the extensive knowledge from large-scale VLMs into lightweight networks [25, 50]. Methods like RISE [19] employ CLIP's extensive knowledge for domain generalization by aligning student image representations with teacher-generated textual embeddings. However, directly minimizing distances between features from different domains may diverge from VLMs' intrinsic contrastive learning framework, potentially limiting the inherent representational diversity.

In contrast, our proposed **VL2Lite** emphasizes direct knowledge transfer within the same domain, aligning with the foundational principles of contrastive learning in VLMs. This domain-specific strategy leverages data similarities to optimize the distillation process. Empirical studies confirm that this approach not only preserves the relevance of transferred knowledge but also enhances student model performance [50], facilitating the development of efficient lightweight models for specific tasks.

3. Methods

In this section, we introduce the comprehensive methodology of our proposed **VL2Lite** framework, designed to refine lightweight neural networks through task-specific training, visual comprehension, and linguistic understanding. The overall architecture of VL2Lite is depicted in Figure 2.

3.1. Overview

Our VL2Lite framework aims to distill knowledge from large-scale Vision-Language Models (VLMs) into lightweight neural networks to improve their performance on task-specific classification tasks, particularly in fine-grained domains and with limited data availability. The key idea is to leverage the rich multimodal representations of VLMs without incurring significant computational overhead during deployment.

The framework involves training a lightweight student model with guidance from a pre-trained VLM, using both visual and linguistic knowledge distillation techniques. During training, the VLM's parameters are kept frozen to preserve their learned representations.

3.2. Task-Specific Classification

To ensure that our lightweight model is effectively tailored to the classification tasks at hand, we define the task-specific classification loss $\mathcal{L}_{\text{cls}}$. This loss function is computed using the cross-entropy between the predicted probabilities and the ground truth labels.

For an input image x and its corresponding ground truth label y, the lightweight model processes x through its encoder to extract feature representations. These features are then passed through a fully connected layer with a softmax activation to produce class probabilities.

The task-specific classification loss for a single observation is mathematically defined as:

$$\mathcal{L}_{\text{cls}} = -\sum_{c=1}^{N_{\text{cls}}} y_{o,c} \log(p(y_{o,c}|x)), \tag{1}$$

where: - N_{cls} is the number of classes, - $y_{o,c}$ is a binary indicator for the correctness of class c for observation o (i.e., $y_{o,c} = 1$ if the ground truth class is c, and $y_{o,c} = 0$ otherwise), - $p(y_{o,c}|x)$ is the predicted probability of class c given input x.

This loss encourages the lightweight model to produce high probabilities for the correct classes, improving its classification accuracy on the task-specific datasets.

3.3. Visual Knowledge Distillation

The objective of the visual knowledge distillation stage is to effectively transfer the visual representation capabilities from the VLM to the lightweight model within the constraints of its reduced complexity. This is achieved through a dimensional reduction mechanism known as the **condensation layer**, which serves as the knowledge projection layer. Additionally, the preservation of relational feature integrity is ensured by employing a dedicated loss function called Relational Knowledge Distillation (RKD) [35].

The condensation layer, designated as $\mathbf{cd}_{\text{img}}(\cdot)$, converts the VLM's high-dimensional feature outputs into a lower-dimensional space suitable for the lightweight model, while the original high-dimensional features from the VLM remain static due to the encoder being frozen. This step is crucial to ensure that the distilled model retains the essential visual information for task-specific classification:

$$\mathbf{I}'_{\text{vlm}} = \mathbf{cd}_{\text{img}}(\text{VLM}_{\text{img}}(x)), \tag{2}$$

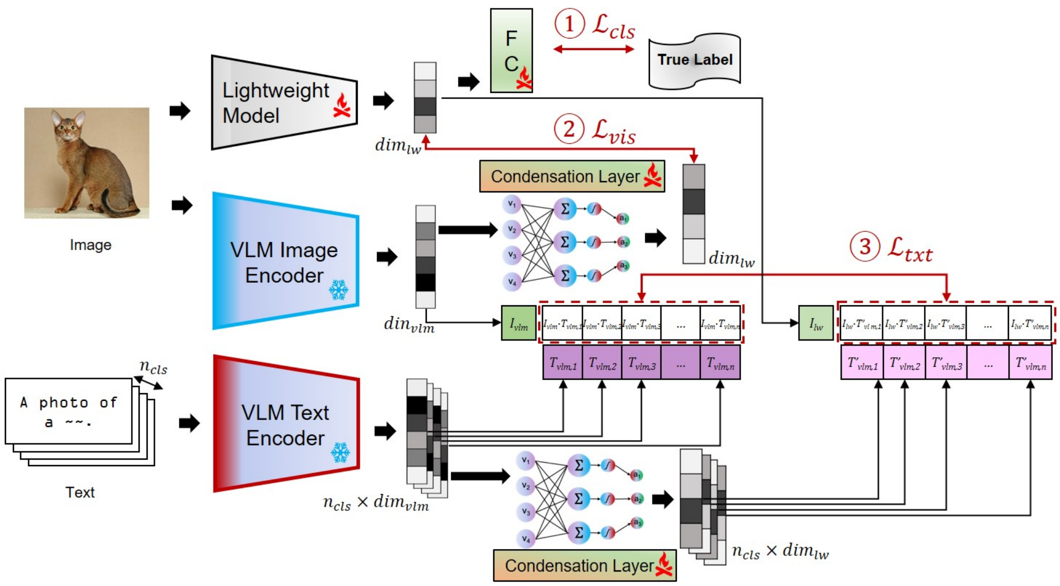

Figure 2. A detailed architecture of the proposed **VL2Lite** framework. The framework comprises three key components: (1) a lightweight model that learns task-specific classification ($\mathcal{L}_{\text{cls}}$), (2) a visual knowledge distillation component ($\mathcal{L}_{\text{vis}}$) that aligns the image representations of the lightweight model with those of the VLM, and (3) a linguistic knowledge distillation component ($\mathcal{L}_{\text{txt}}$) that harmonizes the lightweight model's representations with the VLM's text encoder outputs.

where $\mathbf{I}'_{\text{vlm}}$ represents the transformed feature vector of the VLM after the condensation projection.

Similarly, the lightweight model produces its own feature representation:

$$\mathbf{I}_{\text{lw}} = \text{LW}(x), \tag{3}$$

where $\text{LW}(x)$ is the output of the lightweight model's encoder for input x.

We compute the pairwise Euclidean distances d_{lw} in the lightweight model's feature space as:

$$d_{\text{lw}}(i,j) = \left\| \mathbf{I}_{\text{lw}}(x_i) - \mathbf{I}_{\text{lw}}(x_j) \right\|, \tag{4}$$

This metric reflects the spatial relationship between features as captured by the lightweight model.

Similarly, the pairwise Euclidean distances d'_{vlm} between the transformed VLM feature vectors are:

$$d'_{\text{vlm}}(i,j) = \left\| \mathbf{I}'_{\text{vlm}}(x_i) - \mathbf{I}'_{\text{vlm}}(x_j) \right\|, \tag{5}$$

The visual knowledge distillation loss $\mathcal{L}_{\text{vis}}$ is defined as the sum of discrepancies between the pairwise distances of the lightweight model and those of the transformed VLM features, using the smooth L1 loss function for robustness against outliers:

$$\mathcal{L}_{\text{vis}} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} L1_{\text{smooth}}(d_{\text{lw}}(i,j) - d'_{\text{vlm}}(i,j)), \tag{6}$$

where the smooth L1 loss function $L1_{\text{smooth}}(\delta)$ is defined as:

$$L1_{\text{smooth}}(\delta) = \begin{cases} 0.5\delta^2, & \text{if } |\delta| < 1, \\ |\delta| - 0.5, & \text{otherwise.} \end{cases} \tag{7}$$

By minimizing $\mathcal{L}_{\text{vis}}$, we align the lightweight model's feature space with the relational knowledge structure of the VLM, enabling it to perform complex visual task interpretation efficiently.

3.4. Linguistic Knowledge Distillation

The objective of linguistic knowledge distillation is to enable the lightweight model to comprehend and process textual data, as encoded by the frozen VLM text encoder. This is achieved by aligning the feature space of the lightweight model with that of the VLM using a condensation layer $\mathbf{cd}_{\text{txt}}(\cdot)$.

The VLM's text encoder is frozen to maintain the integrity of its pre-trained language representations. To facilitate compatibility with the lightweight model, we employ the condensation layer:

$$\mathbf{T}_{\text{vlm}} = \text{VLM}_{\text{txt}}(\text{prompt}(y)), \tag{8}$$

$$\mathbf{T}'_{\text{vlm}} = \mathbf{cd}_{\text{txt}}(\mathbf{T}_{\text{vlm}}), \tag{9}$$

where $\text{prompt}(y)$ converts class labels into text prompts (e.g., "A photo of a *[class name]*").

We compute the cosine similarities between the image features and the textual features:

$$\cos_{\text{vlm}} = \frac{\mathbf{I}_{\text{vlm}}^{\top}\mathbf{T}_{\text{vlm}}}{\|\mathbf{I}_{\text{vlm}}\|\|\mathbf{T}_{\text{vlm}}\|}, \tag{10}$$

$$\cos_{\text{lw}} = \frac{\mathbf{I}_{\text{lw}}^{\top}\mathbf{T}'_{\text{vlm}}}{\|\mathbf{I}_{\text{lw}}\|\|\mathbf{T}'_{\text{vlm}}\|}, \tag{11}$$

The linguistic knowledge distillation loss function is formulated as:

$$\mathcal{L}_{\text{txt}} = T^2 \cdot \text{KL}\left(\text{softmax}\left(\frac{\cos_{\text{lw}}}{T}\right), \text{softmax}\left(\frac{\cos_{\text{vlm}}}{T}\right)\right), \tag{12}$$

where T is the temperature parameter that softens the probability distributions, facilitating the distillation of complex linguistic patterns into the lightweight model.

To clarify, each dot product $\cos_{\text{vlm}}$ or $\cos_{\text{lw}}$ is computed *per class/prompt*, resulting in a vector of length N_{cls}. We then apply softmax over this vector to obtain a distribution, which ensures alignment between teacher and student across classes (Eq. (12)).

By minimizing $\mathcal{L}_{\text{txt}}$, the lightweight model learns to align its visual features with the VLM's textual representations, enhancing its semantic understanding.

3.5. Overall Objective Function

The total loss $\mathcal{L}_{\text{total}}$ for the training phase is defined as:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{vis}}\mathcal{L}_{\text{vis}} + \lambda_{\text{txt}}\mathcal{L}_{\text{txt}}, \tag{13}$$

subject to $\lambda_{\text{cls}} + \lambda_{\text{vis}} + \lambda_{\text{txt}} = 1$, where λ_{cls}, λ_{vis}, and λ_{txt} are hyperparameters controlling the relative importance of each loss component.

In our experiments, we initialized $\lambda_{\text{cls}} = 0.01$, and $\lambda_{\text{vis}} = \lambda_{\text{txt}} = 0.495$. During training, λ_{cls} gradually increases, amplifying the importance of the classification loss. This dynamic weighting strategy ensures an initial focus on distilling knowledge from the VLM, followed by a refined emphasis on task-specific accuracy as the model evolves.

The training process involves backpropagating the gradients of $\mathcal{L}_{\text{total}}$ through both the lightweight model and the condensation layers while keeping the VLM's parameters static. This ensures that the distilled knowledge is effectively embedded within the lightweight model's parameters.

3.6. Implementation Details

Vision-Language Model (VLM): We employed the ConvNeXt-XXLarge network [30] from OpenCLIP [2], which provides a 1024-dimensional feature output. For transformer-based lightweight models like Swin Transformer-Tiny, we used the ViT-Large model [4] to maintain architectural compatibility.

Lightweight Models: We used architectures such as ResNet-18 [14], ResNet-34 [14], EfficientNet-B0 [41], MobileNet-V2 [39], ShuffleNet-V2 [31], and Swin Transformer-Tiny [29]. These models were initialized with ImageNet-pretrained weights. Additionally, we included models pre-trained with self-supervised learning methods, such as SEED [6] and DisCo [9].

Condensation Layers: For knowledge projection, we used dual condensation layers. Each condensation layer consists of two fully connected layers with activation functions for nonlinear dimension reduction. Through experimentation, we found that a two-layer MLP provided the best balance between complexity and performance. The layers are not pre-trained and are randomly initialized (Xavier).

Training Details: Models were trained using stochastic gradient descent with momentum. The batch size was set to 256. The temperature parameter T was set to 2. Data augmentation techniques consistent with those used in the experiments were applied.

Self-Supervised Pre-training: VL2Lite was applied to models pre-trained with SEED and DisCo to enhance performance.

Training with Limited Data: To assess learning efficiency under data shortages, models were trained with 30%, 50%, and 100% of the training data.

Code availability. The official implementation is available at https://github.com/jsjangAI/VL2Lite.

3.7. Training Procedure

The training procedure is as follows:

1. **Initialization:** Initialize the lightweight model and condensation layers.
2. **Freeze VLM Parameters:** Keep the VLM's parameters frozen.
3. **Dynamic Weighting:** Set initial values of λ_{cls}, λ_{vis}, and λ_{txt}, adjusting them dynamically during training.
4. **Training Loop:** For each epoch: - Compute $\mathcal{L}_{\text{cls}}$, $\mathcal{L}_{\text{vis}}$, and $\mathcal{L}_{\text{txt}}$. - Update the lightweight model and condensation layers. - Adjust the weighting hyperparameters.
5. **Inference:** After training, deploy only the lightweight model and condensation layers.

3.8. Hardware and Computational Resources

All experiments were conducted on machines equipped with four NVIDIA A4000 GPUs. The models were trained with a batch size of 256, allowing efficient use of GPU memory and reducing training times, particularly when handling large-scale VLMs and performing knowledge distillation across multiple lightweight architectures.

4. Experiments

4.1. Experimental Datasets

For our model's validation, we selected a range of datasets that challenge both task-specific and general classification skills. These datasets include CUB-200-2011 [43] for bird species categorization, FGVC-Aircraft [33] for identifying aircraft models, and NABirds [42] for species recognition. The DTD dataset [3] tests texture recognition abilities, while Oxford-IIIT Pet [36] and Stanford Dogs [21] provide a basis for breed identification in pets and dogs, respectively. The Stanford Cars dataset [22] requires detailed distinctions among car models. Our primary objective is task-specific, fine-grained classification tasks, which reflect a significant demand in practical applications, particularly when dealing with small-scale data. To assess the model's general classification process, we incorporated the well-known Caltech-101 [7] and Caltech-256 [11] datasets, which offer a broad array of object recognition challenges. We adhered to established dataset splits to maintain consistency with conventional studies.

4.2. Comparisons Studies

Comparisons with models w/o KD. The data presented in Table 1 demonstrate a consistent increase in the accuracy of lightweight models through the application of KD from VLMs. Every model, without exception, showed performance improvements, indicating the successful integration of knowledge from VLMs. For "w/o KD," we set both λ_{vis} and λ_{txt} to zero, keeping all other hyperparameters (optimizer, batch size, etc.) identical to the KD setting. Hence, any performance difference is attributed purely to the distillation process.

The improvement is more pronounced in simpler models like ResNet-18 and MobileNet-V2, which benefitted significantly from KD. For instance, ResNet-18 showed a 6.43% increase in accuracy on the CUB-200-2011 dataset, as detailed in Table 1. These results suggest that KD is particularly beneficial for tasks requiring fine-grained feature differentiation.

On the other hand, the advanced Swin Transformer model, despite its inherent complexity and strong learning capability, also demonstrated noticeable improvements through our proposed VL2Lite method. These results highlight the versatility and effectiveness of VL2Lite, underscoring its ability to enhance even the most sophisticated and high-performing models.

Task-specific datasets showed larger gains from KD, reflecting their direct relevance to the distilled knowledge. General-purpose tasks also improved moderately, confirming the broader utility of KD.

In conclusion, the evaluation evidenced by Table 1 shows the efficacy of KD in improving the performance of lightweight models across a range of tasks. The consistent improvements noted across various models and datasets validate KD's utility as a tool for optimizing various lightweight models' accuracy.

Comparisons with other methods. The evaluation presented in Table 2 shows the effectiveness of VL2Lite in enhancing the performance of lightweight models through knowledge distillation. By comparing VL2Lite with traditional KD methods using teacher model as ResNet152 and CLIP, it is evident that VL2Lite achieves superior performance, underscoring its efficacy in distilling complex visual and linguistic knowledge into more compact models.

Significantly, when contrasting VL2Lite with recent approaches like RISE and BorLan, VL2Lite exhibits notable improvements. This is particularly remarkable against the backdrop of RISE, which was among the first to directly distill knowledge from large-scale VLMs for domain generalization. VL2Lite's approach, which focuses on within-domain feature alignment and leveraging contrastive learning principles similar to those in CLIP, demonstrates a more refined strategy. This method ensures a more effective transfer of knowledge, crucial for maintaining the integrity of learned representations in lightweight models.

Furthermore, RISE directly minimizes the distance between embedding features from different domains, a process that recent empirical studies [32] in VLM distillation have suggested might not be as effective as maintaining or contrasting similarities within the same domain. In contrast, our proposed VL2Lite conforms to this latest understanding, optimizing distances more effectively within the same domain and utilizing contrastive learning principles used in CLIP [37] for managing cross-domain variance.

4.3. Ablation Studies

The efficacy of different loss components. In our ablation study, as detailed in Table 3, we studied the effect of various loss functions on the performance of a ResNet-18 model across various datasets designed for task-specific image classification. Adding visual knowledge distillation loss $\mathcal{L}_{\text{vis}}$ has led to some improvements, especially in datasets that require attention to task-specific details, such as CUB-200-2011 and FGVC Aircraft. This suggests that visual KD is important for the model to effectively utilize the fine visual features that differentiate classes.

The results demonstrate that the performance gains derived from linguistic knowledge distillation loss are notably more significant than visual distillation loss. This indicates that the knowledge distilled from the language encoder of VLMs is rich in semantic and structured information. Such text-based knowledge offers robust supervision, effectively guiding the learning process of lightweight models. Consequently, this leads to enhanced generalization capabilities, emphasizing the importance of linguistic knowledge in the distillation process.

Using all three losses together has consistently increased accuracy across all datasets, confirming that a multi-component loss structure is beneficial. Our proposed integration method ensures the comprehensiveness of the distilled knowledge from the teacher model, encompassing

Table 1. Quantitative performance improvements of VL2Lite on various task-specific classification datasets, comparing models before and after applying our proposed KD method. Performance enhancements are highlighted to emphasize the impact of the VL2Lite approach.

Model	Method	Datasets								
		CUB-200	FGVC	NABirds	DTD	Oxford Pet	Sf Dogs	Sf Cars	Ct-101	Ct-256
R-18	w/o KD	64.95%	50.98%	57.01%	67.25%	85.82%	67.80%	70.03%	90.45%	77.58%
	VL2Lite	**71.38%**	**55.87%**	**63.26%**	**71.78%**	**88.72%**	**72.40%**	**77.09%**	**92.33%**	**79.28%**
	Diff.	+6.43%	+4.89%	+6.25%	+4.53%	+2.90%	+4.60%	+7.06%	+1.88%	+1.70%
R-34	w/o KD	69.64%	58.09%	63.05%	70.99%	89.34%	74.10%	74.75%	92.06%	80.99%
	VL2Lite	**76.20%**	**61.60%**	**68.85%**	**72.91%**	**91.30%**	**77.30%**	**81.83%**	**93.24%**	**83.27%**
	Diff.	+6.56%	+3.51%	+5.35%	+1.92%	+1.96%	+3.20%	+7.08%	+1.18%	+2.28%
Eff-B0	w/o KD	66.89%	48.66%	62.91%	69.94%	87.27%	73.80%	66.60%	92.49%	84.38%
	VL2Lite	**73.87%**	**52.18%**	**69.78%**	**72.13%**	**90.02%**	**77.68%**	**75.00%**	**93.19%**	**85.55%**
	Diff.	+6.98%	+3.52%	+6.87%	+2.19%	+2.75%	+3.88%	+8.40%	+0.70%	+1.17%
Mob-V2	w/o KD	64.13%	48.54%	57.79%	67.07%	85.66%	68.14%	68.52%	89.37%	76.13%
	VL2Lite	**71.02%**	**53.82%**	**64.21%**	**69.95%**	**88.06%**	**73.24%**	**74.99%**	**91.15%**	**78.44%**
	Diff.	+6.89%	+5.28%	+6.42%	+2.88%	+2.40%	+5.10%	+6.47%	+1.78%	+2.32%
Shuf-V2	w/o KD	69.40%	50.49%	63.05%	72.21%	87.21%	72.61%	69.34%	92.22%	81.19%
	VL2Lite	**74.99%**	**54.88%**	**69.78%**	**74.22%**	**89.86%**	**76.59%**	**78.11%**	**93.35%**	**84.04%**
	Diff.	+5.59%	+4.39%	+6.71%	+2.01%	+2.65%	+3.98%	+8.77%	+1.13%	+2.86%
Swin-Tiny	w/o KD	80.68%	66.93%	76.11%	77.96%	93.40%	85.20%	81.79%	95.65%	89.28%
	VL2Lite	**83.83%**	**70.93%**	**80.25%**	**79.94%**	**94.52%**	**86.76%**	**86.39%**	**96.14%**	**90.41%**
	Diff.	+3.15%	+4.00%	+3.83%	+2.44%	+1.12%	+1.56%	+4.60%	+0.49%	+1.13%

Table 2. Comparative analysis of classification performance on various datasets using ResNet-18 and MobileNet-V2 models with various KD methods. While most of the KD methods enhance performance, VL2Lite excels by leveraging domain-specific distillation and contrastive learning for improved cross-domain variance management.

Model	Method	Datasets								
		CUB-200	FGVC	NABirds	DTD	Oxford Pet	Sf Dogs	Sf Cars	Ct-101	Ct-256
R-18	w/o KD	64.95%	50.98%	57.01%	67.25%	85.82%	67.80%	70.03%	90.45%	77.58%
	KD (T: R-152)	65.43%	51.94%	57.05%	66.38%	85.99%	67.37%	72.07%	91.90%	77.86%
	RKD (T: R-152)	66.49%	54.58%	58.87%	70.82%	86.74%	68.80%	72.96%	92.22%	78.59%
	KD (T: CLIP)	70.95%	53.83%	62.22%	66.72%	86.74%	68.80%	74.99%	90.28%	77.43%
	RKD (T: CLIP)	68.31%	50.98%	61.16%	68.80%	86.94%	69.03%	72.19%	90.93%	77.98%
	RISE	69.69%	54.81%	59.00%	67.33%	86.81%	68.72%	72.45%	91.68%	78.12%
	BorLan	66.79%	51.14%	59.94%	68.64%	87.28%	69.25%	70.71%	90.50%	78.64%
	VL2Lite	**71.38%**	**55.87%**	**63.26%**	**71.78%**	**88.72%**	**72.40%**	**77.09%**	**92.33%**	**79.28%**
Mob-V2	w/o KD	64.13%	48.54%	57.79%	67.07%	85.66%	68.14%	68.52%	89.37%	76.13%
	KD (T: R-152)	63.48%	49.52%	57.41%	68.12%	86.39%	69.02%	69.31%	89.97%	77.31%
	RKD (T: R-152)	66.09%	52.39%	59.23%	69.47%	86.48%	69.55%	70.89%	90.50%	77.46%
	KD (T: CLIP)	68.00%	48.78%	63.81%	67.90%	86.07%	73.02%	68.74%	88.02%	76.60%
	RKD (T: CLIP)	67.95%	50.29%	60.93%	66.81%	86.56%	69.45%	71.23%	89.48%	76.94%
	RISE	67.51%	52.44%	60.43%	66.64%	86.03%	69.28%	72.67%	89.70%	76.25%
	BorLan	67.38%	52.69%	62.16%	69.60%	87.58%	70.62%	69.08%	90.50%	**78.82%**
	VL2Lite	**71.02%**	**53.82%**	**64.21%**	**69.95%**	**88.06%**	**73.24%**	**74.99%**	**91.15%**	78.44%

Table 3. The efficacy of different loss components in our proposed KD from VLM to ResNet-18 across various datasets

Loss			Dataset					
L_{cls}	L_{vis}	L_{txt}	CUB-200	FGVC	DTD	Oxfords	Sf Dogs	Sf Cars
✓			64.95%	50.98%	67.25%	85.82%	67.80%	70.03%
✓	✓		68.50%	53.31%	68.46%	87.32%	68.85%	71.86%
✓		✓	71.03%	54.84%	69.42%	87.98%	71.86%	76.68%
✓	✓	✓	**71.38%**	**55.87%**	**71.78%**	**88.72%**	**72.40%**	**77.09%**

visual, semantic, and class-related information.

In conclusion, our study supports the importance of combining different loss functions in KD. The collaborative impact of $\mathcal{L}_{\text{cls}}$, $\mathcal{L}_{\text{vis}}$, and $\mathcal{L}_{\text{txt}}$ provides the lightweight models with a richer understanding, enabling them to make more accurate predictions. This approach of integrating multiple forms of knowledge into a single distillation process enhances the student model's learning, offering a robust method for knowledge transfer and model refinement.

The number of layers in condensation projection. In our additional ablation study, we investigated the impact of varying the number of layers in condensation projection on

Table 4. Comparative analysis of the effect of the different number of knowledge projection layer configurations on the classification performance using ResNet-18 across various datasets.

# of layer	Dataset					
	CUB-200	FGVC	DTD	Oxfords	Sf Dogs	Sf Cars
Single layer	71.38%	55.65%	70.12%	88.28%	72.36%	**77.22%**
Multi 2 layers	71.38%	**55.87%**	**71.78%**	**88.72%**	**72.40%**	77.09%
Multi 3 layers	**71.57%**	55.78%	70.55%	88.36%	72.22%	76.10%

Table 5. Performance comparison of ResNet-18 models pre-trained using self-supervised learning methods SEED [6] and DisCo [9], with and without additional KD from VLMs

Model	Method	Datasets				
		CUB-200	DTD	Oxford Pet	Sf Dogs	Sf Cars
SEED	w/o KD	63.16%	72.12%	84.51%	66.30%	73.24%
	VL2Lite	**69.99%**	**74.13%**	**87.24%**	**70.78%**	**77.85%**
	Diff.	+6.83%	+2.01%	+2.73%	+4.48%	+4.61%
DisCo	w/o KD	59.13%	68.38%	79.88%	62.10%	70.19%
	VL2Lite	**64.19%**	**72.21%**	**84.00%**	**68.71%**	**71.23%**
	Diff.	+5.06%	+3.83%	+4.12%	+6.61%	+1.04%

the classification performance. As indicated in Table 4, the number of layers does not significantly affect the KD efficacy. However, a two-layer MLP condensation was observed to yield the best performance across the majority of datasets. Consequently, the two-layer MLP was adopted for its optimal balance of complexity and performance.

KD to Self-supervised Pre-trained Weights Our study extends the capabilities of ResNet-18 models pre-trained with self-supervised learning techniques, SEED [6] and DisCo [9]. These techniques initially adapt the architecture from a ResNet-50x2 backbone to a streamlined ResNet-18 framework, setting a strong foundation for knowledge assimilation. By integrating KD from VLMs, we sought to enhance the pre-trained networks' representational capacity. As indicated in Table 5, integrating VLM-based distillation with SSL pre-training not only reinforces the efficacy of the SSL models but also has the adaptability of these models to additional knowledge enhancement. This strategy of enriching SSL pre-trained weights with VLM knowledge exhibits consistent performance boosts across diverse datasets. It emphasizes the importance: both SSL and other pre-training paradigms benefit from the infusion of VLM knowledge, suggesting an effective approach for advancing lightweight model performance. This result suggests that the proposed method could be a standard strategy for improving various pre-trained models, marking a substantial step forward in the development of efficient models used for practical applications.

Learning Efficiency in data shortages. The quantity of training data is important factor that significantly influences model performance. However, in practical applications, there can be a possibility of facing data scarcity which can cause low performance. To address this, our research presents some evaluations of the performance of the proposed method with limited training data, as shown in Table 6. We assessed the impact of training ResNet-18 with only 30%, 50%, and the full extent of training data, implementing VL2Lite to evaluate its robustness in data-deficient environments. The results indicate that our proposed VL2Lite substantially enhances task-specific performance even when the available training data is deficient. For example, with 30% of the training data, VL2Lite-enhanced ResNet-18 backbone shows over a 10% increase in accuracy on the CUB-200 dataset. These improvements with reduced data are consistent across various datasets, with the Stanford Cars dataset experiencing an increase of more than 10% in accuracy.

Table 6. Quantitative analysis of ResNet-18 with 30%, 50% and 100% training data, showing enhanced performance through our proposed VL2Lite across various datasets.

Ratio	Method	Datasets					
		CUB-200	FGVC	DTD	Oxford Pet	Sf Dogs	Sf Cars
30%	w/o KD	39.61%	26.82%	58.88%	59.75%	59.75%	32.84%
	VL2Lite	**50.32%**	**33.72%**	**61.15%**	**83.43%**	**65.41%**	**43.36%**
	Diff.	+10.71%	+6.90%	+2.27%	+3.58%	+5.66%	+10.52%
50%	w/o KD	51.22%	39.33%	60.88%	80.21%	63.54%	50.06%
	VL2Lite	**60.87%**	**45.18%**	**62.63%**	**84.93%**	**69.13%**	**60.41%**
	Diff.	+9.65%	+5.25%	+1.75%	+4.72%	+5.59%	+10.35%
100%	w/o KD	64.95%	50.98%	67.25%	85.82%	67.80%	70.03%
	VL2Lite	**71.38%**	**55.87%**	**71.78%**	**88.72%**	**72.40%**	**77.09%**
	Diff.	+6.43%	+4.89%	+4.53%	+2.90%	+4.60%	+7.06%

Table 7. Comparative Performance with other visual distillation loss: L2 Norm, Cross-modal loss. The backbone is ResNet-18

Method	Datasets					
	CUB-200	FGVC	DTD	Oxford Pet	Sf Dogs	Sf Cars
L2 Norm	71.35%	**56.80%**	69.60%	88.38%	72.12%	76.57%
Cross-Modal loss	67.43%	53.92%	69.69%	86.56%	69.52%	71.45%
VL2Lite	**71.38%**	55.87%	**71.78%**	**88.72%**	**72.40%**	**77.09%**

These improvements can show the efficiency of VL2Lite in knowledge retention and transfer from VLMs, allowing the model to maintain high performance with less data. The adaptability of VL2Lite to various data scales presents its potential as a solution for scenarios with limited data availability, positioning it as a beneficial approach for maintaining model efficacy in data-constrained environments.

Other loss studies. Regarding the use of visual distillation loss such as L2 Norm minimization, we evaluated its impact in comparison to our proposed L1-based approach. Our empirical results, summarized in Table 7, indicate that while L2 loss performs comparably, VL2Lite generally achieves higher accuracy across multiple datasets. Due to the robustness of L1 loss against outliers, which might be more prevalent in the visual features represented in high-dimensional space. Furthemore, we diverge from RISE by focusing on image–text cross-modality relationships. Table 7 shows our method achieving a performance boost of 1.9% to 5.6% over cross-modal loss across various datasets. This indicates that utilizing similarities within the same modality can be more effective than direct image-text knowledge transfer, highlighting a simpler yet powerful approach.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 1, 2

[2] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. 1, 2, 5

[3] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 6

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 5

[5] Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lijuan Wang, Yezhou Yang, and Zicheng Liu. Compressing visual-linguistic model via knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1428–1438, 2021. 3

[6] Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, and Zicheng Liu. Seed: Self-supervised distillation for visual representation. *arXiv preprint arXiv:2101.04731*, 2021. 2, 3, 5, 8

[7] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 6

[8] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023. 1, 2

[9] Yuting Gao, Jia-Xin Zhuang, Shaohui Lin, Hao Cheng, Xing Sun, Ke Li, and Chunhua Shen. Disco: Remedy self-supervised learning on lightweight models with distilled contrastive learning. *arXiv preprint arXiv:2104.09124*, 2021. 1, 2, 3, 5, 8

[10] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819, 2021. 1

[11] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007. 6

[12] Jindong Gu, Ahmad Beirami, Xuezhi Wang, Alex Beutel, Philip Torr, and Yao Qin. Towards robust prompts on vision-language models. *arXiv preprint arXiv:2304.08479*, 2023. 2

[13] Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*, 2023. 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[15] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3779–3787, 2019. 1

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2

[17] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2

[19] Zeyi Huang, Andy Zhou, Zijian Ling, Mu Cai, Haohan Wang, and Yong Jae Lee. A sentence speaks a thousand images: Domain generalization through distilling clip with language guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11685–11695, 2023. 3

[20] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024. 2

[21] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*. Citeseer, 2011. 6

[22] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 6

[23] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 1

[24] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 1, 2

[25] Xuanlin Li, Yunhao Fang, Minghua Liu, Zhan Ling, Zhuowen Tu, and Hao Su. Distilling large vision-language model with out-of-distribution generalizability. In *Proceed-*

ings of the IEEE/CVF International Conference on Computer Vision, pages 2492–2503, 2023. 3

[26] Xin Li, Dongze Lian, Zhihe Lu, Jiawang Bai, Zhibo Chen, and Xinchao Wang. Graphadapter: Tuning vision-language models with dual knowledge graph. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2

[28] Yufan Liu, Jiajiong Cao, Bing Li, Chunfeng Yuan, Weiming Hu, Yangxi Li, and Yunqiang Duan. Knowledge distillation via instance relationship graph. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7096–7104, 2019. 1

[29] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 5

[30] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 5

[31] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018. 5

[32] Zongyang Ma, Guan Luo, Jin Gao, Liang Li, Yuxin Chen, Shaoru Wang, Congxuan Zhang, and Weiming Hu. Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14074–14083, 2022. 1, 6

[33] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 6

[34] Utkarsh Ojha, Yuheng Li, Anirudh Sundara Rajan, Yingyu Liang, and Yong Jae Lee. What knowledge gets distilled in knowledge distillation? *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2

[35] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3967–3976, 2019. 2, 3

[36] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 6

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 6

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[39] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 5

[40] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 1, 2

[41] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *In International Conference on Machine Learning*, page 6105–6114, 2019. 5

[42] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 595–604, 2015. 6

[43] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 5

[44] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022. 1, 2

[45] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19175–19186, 2023. 1, 2

[46] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022. 2

[47] Yuzhu Wang, Lechao Cheng, Manni Duan, Yongheng Wang, Zunlei Feng, and Shu Kong. Improving knowledge distillation via regularizing feature norm and direction. *arXiv preprint arXiv:2305.17007*, 2023. 2

[48] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. 2013. 1, 2

[49] Zhecan Wang, Noel Codella, Yen-Chun Chen, Luowei Zhou, Xiyang Dai, Bin Xiao, Jianwei Yang, Haoxuan You, Kai-Wei Chang, Shih-fu Chang, et al. Multimodal adaptive distillation for leveraging unimodal encoders for vision-language tasks. *arXiv preprint arXiv:2204.10496*, 2022. 3

[50] Chuanguang Yang, Zhulin An, Libo Huang, Junyu Bi, Xinqiang Yu, Han Yang, and Yongjun Xu. Clip-kd: An

empirical study of distilling clip models. *arXiv preprint arXiv:2307.12732*, 2023. 1, 3

[51] Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10899–10909, 2023. 2

[52] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 1, 2

[53] Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6995–7004, 2021. 2