ComKD-CLIP: Comprehensive Knowledge Distillation for Contrastive Language-Image Pre-traning Model

Yifan Chen[1,3*], Xiaozhen Qiao[2,3*], Zhe Sun[1,3†] & Xuelong Li[1,3‡]

[1]School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University
[2]School of Information Science and Technology, University of Science and Technology of China
[3]Institute of Artificial Intelligence (TeleAI), China Telecom

Abstract

Contrastive Language-Image Pre-training (CLIP) models excel in integrating semantic information between images and texts through contrastive learning techniques. It has achieved remarkable performance in various multimodal tasks. However, the deployment of large CLIP models is hindered in resource-limited environments, while smaller models frequently fail to meet the performance benchmarks required for practical applications. In this paper, we propose a novel approach, ComKD-CLIP: Comprehensive Knowledge Distillation for Contrastive Language-Image Pre-traning Model, which aims to comprehensively distill the knowledge from a large teacher CLIP model into a smaller student model, ensuring comparable performance with significantly reduced parameters. ComKD-CLIP is composed of two key mechanisms: Image Feature Alignment (IFAlign) and Educational Attention (EduAttention). IFAlign makes the image features extracted by the student model closely match those extracted by the teacher model, enabling the student to learn teacher's knowledge of extracting image features. EduAttention explores the cross-relationship between text features extracted by the teacher model and image features extracted by the student model, enabling the student model to learn how the teacher model integrates text-image features. In addition, ComKD-CLIP can refine the knowledge distilled from IFAlign and EduAttention by leveraging the text-image feature fusion results of the teacher model, ensuring the student model accurately absorbs the teacher's knowledge. Extensive experiments conducted on 11 datasets have demonstrated the superiority of the proposed method.

Introduction

Contrastive Language-Image Pre-training (CLIP) (Radford et al. 2021) has emerged as a prominent pre-trained Visual Language Model (VLM) due to its ability to effectively integrate the semantic information between images and texts through contrastive learning techniques. This ability has enabled CLIP to demonstrate strong performance across various multimodal tasks such as image recognition, visual question answering, and image description generation (Li et al. 2023a; Wei et al. 2022; Singh et al. 2022). However, large CLIP models cannot be deployed in resource-constrained environments and the performance of smaller CLIP models often falls short of application requirements. These limitations significantly restrict the practical applicability of CLIP models.

Knowledge Distillation (KD) (Hinton, Vinyals, and Dean 2015) has been introduced into CLIP to address these issues. CLIP-TD (Wang et al. 2022) distills knowledge from both

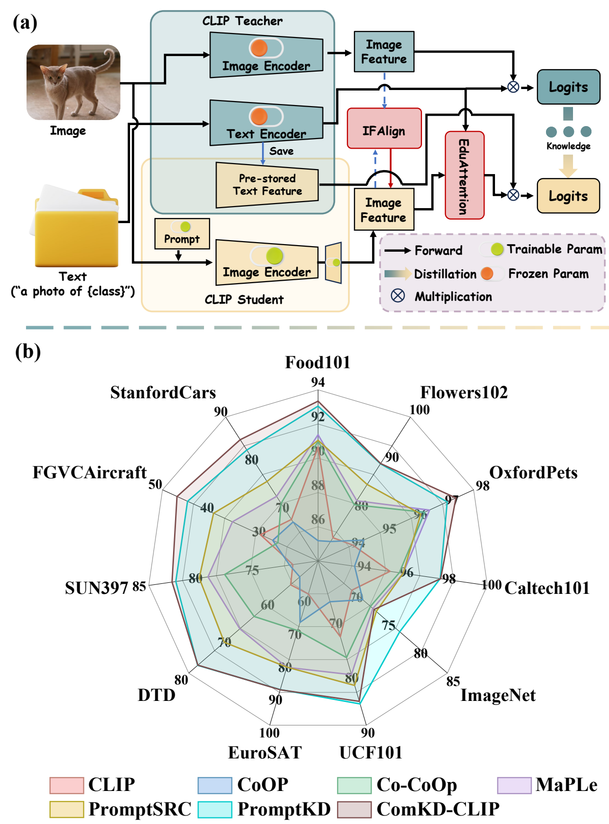

Figure 1: (a) Schematic of proposed ComKD-CLIP framework. (b) Comparison of the Harmonic Mean (HM) for generalization from base to novel categories. All methods utilize the ViT-B/16 image encoder from the pretrained CLIP model. ComKD-CLIP is proud to achieve the **best performance** in 8 out of 11 diverse recognition datasets.

*These authors contributed equally.
†Corresponding author: sunzhe@nwpu.edu.cn
‡Corresponding author: li@nwpu.edu.cn

CLIP's vision and language branches into existing architectures for VL tasks. TinyCLIP (Wu et al. 2023) proposes affinity mimicking and weight inheritance to improve the performance of small models by leveraging large-scale models. CLIPPING (Pei et al. 2023) proposes a new layer-wise alignment with the student as the base, which enables the student to fully absorb the knowledge of the teacher. PromptKD (Li et al. 2024b) uses unlabeled domain data to perform prompt-based knowledge distillation for CLIP, which can greatly improve the performance of small model. CLIP-KD (Yang et al. 2024) proposes several distillation strategies to examine the effectiveness of CLIP-KD. However, these methods distill the knowledge of the teacher model based on the results of text-image feature fusion, ignoring the knowledge embedded within the fusion process. This oversight substantially hinders the student model to absorb the teacher model's knowledge.

To ensure that the student model can fully absorb the knowledge of the teacher model, we endeavor to distill the knowledge embedded in the feature fusion process of teacher model and refine the distilled knowledge leveraging the feature fusion results of teacher model. Accordingly, we propose ComKD-CLIP: Comprehensive Knowledge Distillation for Contrastive Language-Image Pre-traning Model. ComKD-CLIP is composed of two key modules: Image Feature Alignment (IFAlign) and Educational Attention (EduAttention). During the feature fusion stage: IFAlign ensures that the image features extracted by the student model closely match those extracted by the teacher model. This alignment enable the student to absorb the teacher's knowledge of how to extract image feature. Concurrently, EduAttention explores the cross-relationship between the text features extracted by the teacher model and the image features extracted by the student model. Through this strategy, EduAttention enables the student model to comprehend and emulate the teacher model's abilities for integrating image and text features, thus enriching its own mutlimodal understanding. Furthermore, to make the student model accurately absorb the knowledge of the teacher model, ComKD-CLIP refines the knowledge distilled from ImgAlign and EduAttention leveraging the feature fusion results of the teacher model. The schematic of proposed ComKD-CLIP is shown as Fig. 1 (a). We also compare the performance of the proposed method with some state-of-the-art methods on 11 datasets. The experimental results are shown in Fig. 1 (b). It is worth mentioning that the proposed method has best performance in 8 out of 11 diverse recognition datasets.

The main contributions of the proposed method can be summarized as follows:

- We propose an IFAlign module that enables student model to absorb the teacher model's knowledge on how to extract image features during the process of text-image feature fusion.
- We propose an EduAttention module that enables student model to absorb the teacher model's knowledge on how to integrate the text-image features during the process of text-image feature fusion.
- We refine the knowledge distilled from IFAlign and EduAttention leveraging the feature fusion results of the teacher model, which can prompt the student model to accurately absorb the teacher model's knowledge. Extensive experiments conducted on 11 datasets have demonstrated the superiority of the proposed ComKD-CLIP.

Related Work

Contrastive Language-Image Pre-training (CLIP)

CLIP can simultaneously comprehend and fuse the text-image data, achieving outstanding performance in multimodal tasks (Gao et al. 2022; Zhu et al. 2023; Li et al. 2024a). SLIP (Mu et al. 2022) extends CLIP's capabilities by integrating it with self-supervised learning, facilitating its application within the domain of multi-task learning. MaskCLIP (Dong et al. 2023) innovates further by introducing masked self-distillation, a technique that distills the representation of a complete image onto the predicted representation of its masked counterpart. This approach significantly enhances CLIP's performance. AttCLIP (Yang et al. 2023) incorporates attention mechanisms into CLIP, enabling the model to selectively focus on tokens that exhibit a high degree of correlation with the corresponding textual information. This method not only facilitates efficient multi-view learning but also economizes on training time. CLIP-Decoder (Ali and Khan 2023) enriches the multi-modal representation learning of CLIP through the integration of distinct encoders for text and images, leading to significant advancements in multi-label classification tasks. MoPE-CLIP (Lin et al. 2024) proposes a novel module-wise pruning error metric, allowing for the effective leveraging of teacher model knowledge. This provides a unified solution for the pre-training phase of CLIP models. Collectively, these advancements underscore CLIP's prowess in multimodal tasks. However, the deployment of large CLIP models remains constrained in resource-limited environments, whereas small models often fail to meet the requisite benchmarks for practical utility. Consequently, the primary challenge lies in compressing CLIP models without compromising their performance, thus facilitating their widespread applicability across diverse computational landscapes.

Knowledge Distillation (KD)

KD aims to enable the small student model to absorb knowledge from the large teacher model, thus achieving comparable performance to the large model. It has achieved remarkable success in numerous vision tasks, including image segmentation (Liu et al. 2019; Yang et al. 2022), object detection (Jia et al. 2024; Wang et al. 2024), and pose estimation (Li et al. 2021). Recently, many researchers have endeavored to introduce KD into CLIP, motivated by the pressing need to surmount the operational challenges faced by large CLIP models in resource-limited environments, alongside the performance shortfalls exhibited by smaller models in practical applications.(Laroudie et al. 2023). CLIPPING (Pei et al. 2023) introduces a novel layer-wise alignment strategy that takes the student model as the foundation, enabling the student model to effectively absorb the knowledge from the teacher model. PromptKD (Li et al. 2024b), in a distinct approach, capitalizes on unlabeled domain data to facilitate prompt-based knowledge distillation within the CLIP paradigm, significantly bolstering the performance of smaller CLIP models. TinyCLIP (Wu et al. 2023), similarly targeting CLIP distillation, achieves commendable results through the innovative application of affinity mimicking and weight inheritance techniques. However, prior studies predominantly concentrate on distilling the knowledge of teacher model based on feature fusion results, overlooking the intricate knowledge encapsulated within the feature fusion process. In stark contrast to existing distillation methodologies, the proposed method uniquely distills the knowledge inherent to the text-image feature fusion process within teacher models. In addition, by refining the distilled knowledge leveraging the feature fusion results, ComKD-CLIP enables the student model to accurately absorb the nuanced knowledge from teacher model, thereby precipitating a marked improvement in their performance.

Approach

Preliminaries

CLIP is one of the most commonly used VLMs, comprising independent image and text encoder branches. It aligns and fuses images with texts to learn a joint multimodal embedding space. In the image encoding branch, a labeled visual recognition dataset $D = \{x_j, y_j\}_{j=1}^{M}$ is used as input. Each image x from dataset D is processed by the image encoder f_I to obtain normalized image features $u = f_I(x)/||f_I(x)||_2 \in \mathbb{R}^d$. Corresponding to the image recognition dataset D has N class names c $= \{c_i\}_{i=1}^{N}$. In the text encoder branch, the input data is text descriptions t_i generated from the template "a photo of a $\{c_i\}$". Each t_i, after being encoded by the text encoder, yields normalized text features $w_i = f_T(t_i)/||f_T(t_i)||_2 \in \mathbb{R}^d$, where d is the dimension of the text features. The collection of all text features $W = [w_1, w_2, \ldots, w_N] \in \mathbb{R}^{N\times d}$ serves as the classification weight matrix. Based on this data, the classification output probability can be calculated as follows:

$$p(y|x) = \frac{\exp(uw_y^{\mathsf{T}}/\tau)}{\sum_{i=1}^{N}\exp(uw_i^{\mathsf{T}}/\tau)}, \quad (1)$$

where uw^{T} represents the output logit and τ is the temperature parameter.

KD is originally proposed by Hinton et al. (Hinton, Vinyals, and Dean 2015), it transfers knowledge from a large, pretrained teacher model to a smaller, lightweight student model. It can help the student to absorb the teacher's knowledge for efficient deployment. This process employs the KL divergence loss to align the feature distributions of both models. The KL divergence loss is defined as follows:

$$L_{kd}(q^t, q^s, \tau) = \tau^2 KL(\sigma(q^t/\tau), \sigma(q^s/\tau)), \quad (2)$$

where q^t and q^s represent the logits predicted by the teacher and student models, respectively. $\sigma(\cdot)$ represents the softmax function, and τ is the temperature parameter (Hinton, Vinyals, and Dean 2015; Li et al. 2023b), which adjusts the smoothness of the probability distribution.

Pipline

As illustrated in Fig. 2, our proposed ComKD-CLIP framework comprises two principal stages: the pretraining of the large CLIP teacher model and the subsequent training of a small CLIP student model. In the initial phase, as delineated in Fig. 2(a), the large CLIP teacher model is pretrained on a labeled domain dataset, $D_{labeled} = \{x_i, y_i\}_{i=1}^{M}$, to enhance its performance, aligning with contemporary methodologies such as PromptSRC (Khattak et al. 2023b) and PromptKD (Li et al. 2024b). Innovatively, we incorporate learnable prompts into both the image and text encoder branches of the teacher model via a concatenation strategy. The image and text data from the labeled domain dataset are processed through the image encoder f_V^t and text encoder f_T^t, respectively, producing image features $u_t^p \in \mathbb{R}^d$ and text features $w_t^p \in \mathbb{R}^d$. The ultimate output logits q^t is calculated by Eq. 1. The training of the teacher model entails minimizing the cross-entropy loss between the predicted probability distributions and the true labels, thereby optimizing the model's parameters. This rigorous pretraining phase ensures the teacher model acquires robust knowledge that can be effectively distilled to the student model during the latter stage of our framework.

As depicted in Fig. 2(b), the student CLIP model directly capitalizes on the pre-trained text features from the teacher model, thereby significantly curtailing the training costs with the text encoder branch. Simultaneously, a lightweight CLIP image encoder branch is engineered within the student model to decrease resource costs while maintaining competitive performance. During the processing of input data from the unlabeled domain dataset $D_{unlabeled}$ by the student model's image encode, we incorporate the IFAlign module. This module serves to align the student model's image features $u_s^p \in \mathbb{R}^d$ with the teacher model's image features $u_t^p \in \mathbb{R}^d$, thereby facilitating the student model to absorb the knowledge of how the teacher model extracts salient image features. Subsequent to the feature alignment, the EduAttention module is introduced to explore the cross-relationship between the image features extracted by student model and text features provided by teacher model. This exploration enables the student model to learn the nuanced strategies employed by the teacher model for integrating text-image feature. In addition, we employ the KL divergence to minimize the discrepancy between the logits produced by the teacher and student models. This optimization ensures that the knowledge distilled by the student model is refined and closely mirrors that of the teacher, thus enhancing the student model to accurately absorb the knowledge of teacher model. Finally, the inference process of trained student model is shown in Fig. 2(d).

ComKD-CLIP

IFAlign: The schematic of IFAlign is illustrated in Fig. 2(c). To make the image features extracted by the student model closely match those extracted by the teacher model, we align the mean and variance statistics for extracted features. The calculation process can be formulated as follows:

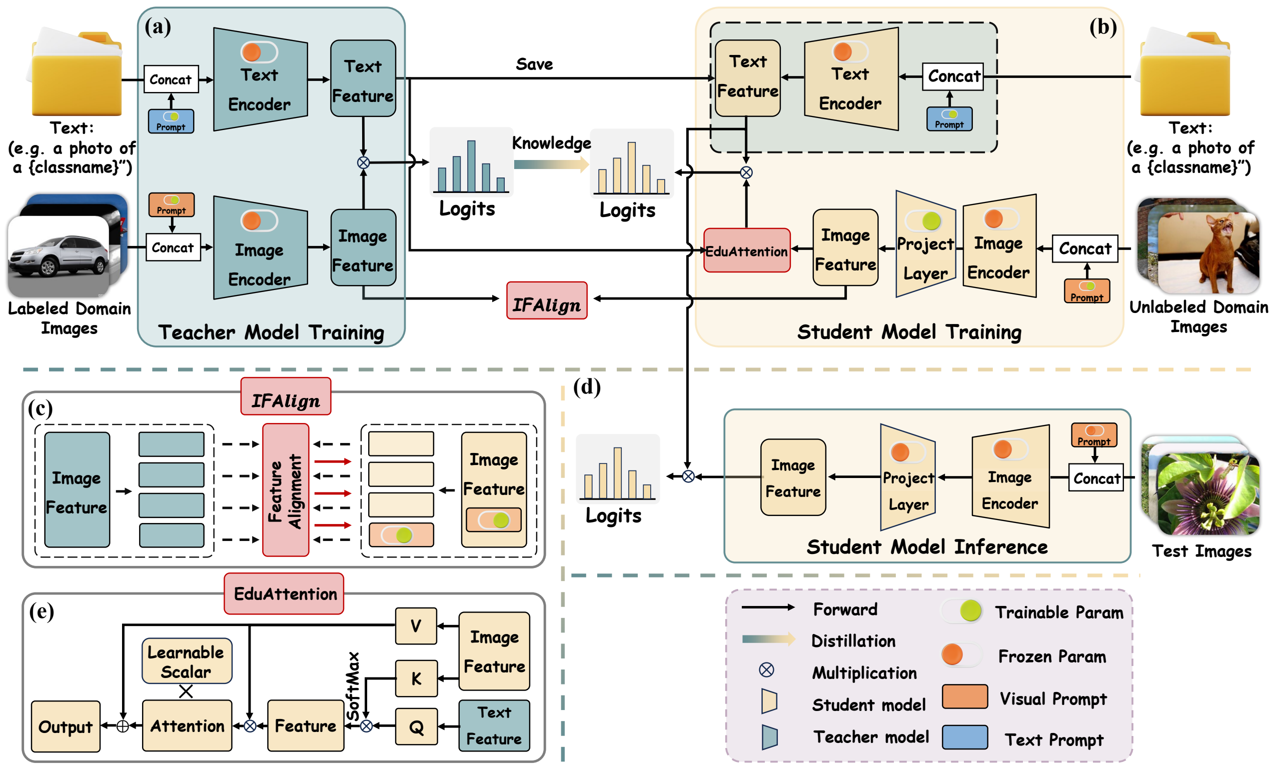

Figure 2: Overview of our proposed ComKD-CLIP framework. (a) Utilization of the cue learning method with a well-trained large teacher CLIP model; (b) A smaller student CLIP model, which is trained with learnable cues and reuses the text features from the teacher model, requiring training only for the image encoder branch; (c) The schematic of IFAlign module; (d) The inference process within the trained student model, where the text encoder branch reuses the text features of the teacher model; (e) The schematic of EduAttention module.

$$\begin{aligned}\boldsymbol{\mu}_{\mathrm{s}}(u_s;\boldsymbol{p}) &= \frac{1}{N}\sum P(\boldsymbol{u}_s^p),\\ \boldsymbol{\mu}_{\mathrm{t}}(u_t;\boldsymbol{p}) &= \frac{1}{N}\sum \boldsymbol{u}_t^p,\end{aligned} \tag{3}$$

$$\begin{aligned}\boldsymbol{\sigma}_s^2(u_s;\boldsymbol{p}) &= \frac{1}{N}\sum\Big(P(\boldsymbol{u}_s^p)-\boldsymbol{\mu}_s(u_s;\boldsymbol{p})\Big)^2,\\ \boldsymbol{\sigma}_t^2(u_t;\boldsymbol{p}) &= \frac{1}{N}\sum\Big(\boldsymbol{u}_t^p-\boldsymbol{\mu}_t(u_t;\boldsymbol{p})\Big)^2,\end{aligned} \tag{4}$$

where $\boldsymbol{\mu}_{\mathrm{s}}(u_{\mathrm{s}};\boldsymbol{p})$ and $\boldsymbol{\sigma}_{\mathrm{s}}^2(u_{\mathrm{s}};\boldsymbol{p})$ represent the means and variances of the image features extracted by the student model; $\boldsymbol{\mu}_{\mathrm{t}}(u_{\mathrm{t}};\boldsymbol{p})$ and $\boldsymbol{\sigma}_{\mathrm{t}}^2(u_{\mathrm{t}};\boldsymbol{p})$ correspond to those extracted by the teacher model. $\boldsymbol{u}_{\mathrm{s}}^p$ and $\boldsymbol{u}_{\mathrm{t}}^p$ denote the image features with prompts for the student and teacher models, respectively. The learnable projector $P(\cdot)$ within the student's image encoder branch is designed to adjust feature dimensions efficiently and cost-effectively, ensuring precise alignment. Following this, we use the L_1 loss to align the mean and variance of the image features extracted by student model with teacher model. This alignment can facilitate the student model to absorb the knowledge of how the teacher model extracts salient image features. The specific alignment process can be formulated as follows:

$$\begin{aligned}\mathcal{L}_{\text{align\_mean}} &= \|\boldsymbol{\mu}_{\mathrm{s}}(u_s;\boldsymbol{p})-\boldsymbol{\mu}_{\mathrm{t}}(u_t;\boldsymbol{p})\|_1,\\ \mathcal{L}_{\text{align\_var}} &= \|\boldsymbol{\sigma}_s^2(u_s;\boldsymbol{p})-\boldsymbol{\sigma}_t^2(u_t;\boldsymbol{p})\|_1,\\ \mathcal{L}_{\text{align}} &= \mathcal{L}_{\text{align\_mean}}+\mathcal{L}_{\text{align\_var}},\end{aligned} \tag{5}$$

where $\mathcal{L}_{\text{align\_mean}}$ represents the difference between the mean values of image features extracted by the teacher model and the student model, $L_{\text{align\_var}}$ represents the difference between the variances values of image features extracted by the teacher model and the student model. Combining $\mathcal{L}_{\text{align\_mean}}$ and $L_{\text{align\_var}}$ as the alignment loss $\mathcal{L}_{\text{align}}$ allows the student model to fully absorb the knowledge of how the teacher model extracts image features.

EduAttention: The schematic of EduAttention is illustrated in Fig. 2(e). In this module, an attention mechanism is leveraged to explore the cross-relationship between the image features extracted by student model and text features provided by teacher model, which can facilitate the student model to learn the nuanced strategies employed by the teacher model for integrating text-image feature. The specific calculation process can be formulated as follows:

$$\begin{aligned}&Q = FC(\boldsymbol{w}_t^p),\quad K = FC(\boldsymbol{u}_s^p),\quad V = FC(\boldsymbol{u}_s^p),\\ &f_{att} = Softmax\left(\frac{QK^T}{\sqrt{C}}\right)\cdot V,\end{aligned} \tag{6}$$

where $\boldsymbol{u}_s^p$ represents the image features extracted by the student models, $\boldsymbol{w}_t^p$ represents the text features extracted by the teacher models, f_{att} represents the cross-relationship between $\boldsymbol{u}_s^p$ and $\boldsymbol{w}_t^p$, C is a hyperparameter, and $FC(\cdot)$ represents the fully connected layer.

To integrate the knowledge absorbed from the teacher model by IFAlign and EduAttention, we multiply f_{att} by a learnable parameter α and perform a element-wise sum operation with the extracted image features text $\boldsymbol{u}_t^p$ to final image features f_e. The specific calculation process can be formulated as follows:

$$f_e = \boldsymbol{u}_s^p + \alpha \cdot f_{att}, \tag{7}$$

where α is initialized as 0 and gradually learns to assign more weight.

Distilled Knowledge Refinement: After the student model absorbs the knowledge of how the teacher model extracts image features and combines text-image features, we try to refine the absorbed knowledge based on the feature fusion results of the teacher model. As depicted in Fig. 2, we utilize the KL divergence to minimize the discrepancy between the feature distribution produced by the teacher and student models. The specific process can be formulated as follows:

$$\mathcal{L}_{stu} = \mathcal{L}_{\text{kd}}(q^t, q^s, \tau), \tag{8}$$

where q^t and q^s represent the logits predicted by the teacher and student models, respectively, which is calculated by the corresponding image features and text features using Eq. 1. τ is the temperature parameter, which is used to adjust the smoothness of the probability distribution.

Finally, we combine the student model's alignment loss $\mathcal{L}_{\text{stu}}$ with the feature distribution loss $\mathcal{L}_{\text{align}}$ as the final loss function to train the parameters of the small CLIP model, the specific loss formula is shown as follows:

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{stu}} + \mathcal{L}_{\text{align}}. \tag{9}$$

Experiments

Settings

Datasets: In this study, we adopt the methodologies from PromptSRC (Khattak et al. 2023b) and PromptKD (Li et al. 2024b) to assess the generalization from a base to novel classes, cross-dataset evaluation. We utilize 11 diverse image recognition datasets encompassing a range of tasks such as:

- Generic object recognition with ImageNet (Deng et al. 2009) and Caltech101 (Fei-Fei, Fergus, and Perona 2004).

- Fine-grained classification using OxfordPets (Parkhi et al. 2012), StanfordCars (Krause et al. 2013), Flowers102 (Nilsback and Zisserman 2008), Food101 (Bossard, Guillaumin, and Van Gool 2014), and FGVCAircraft (Maji et al. 2013).

- Scene recognition with SUN397 (Xiao et al. 2010).

- Action recognition from UCF101 (Soomro, Zamir, and Shah 2012).

- Texture classification via DTD (Cimpoi et al. 2014).

- Satellite imagery with EuroSAT (Helber et al. 2019).

For the domain generalization benchmark, ImageNet (Deng et al. 2009) serves as the source dataset, with ImageNetA (Hendrycks et al. 2021b), ImageNet-R (Hendrycks et al. 2021a), ImageNet-Sketch (Wang et al. 2019), and ImageNetV2 (Recht et al. 2019) as the out-of-distribution test datasets.

Implementation Details: We employ the ViT-L/14 CLIP model as the teacher model and the ViT-B/16 CLIP model as the student model for our ComKD-CLIP framework. Following the PromptKD configuration, we set the prompt depth to 9, with both vision and language prompt lengths fixed at 4. Optimization is carried out using Stochastic Gradient Descent (SGD) with the temperature hyperparameter τ set to its default value of 1. Initial text prompts for the first layer are generated using embeddings of the phrase "a photo of a {classname}". We report the base and novel class accuracies along with their Harmonic Mean (HM), averaged over three runs. All experiments are conducted on a single Nvidia A100 GPU.

Base-to-novel Generalization

Following (Zhou et al. 2022a) (Khattak et al. 2023a) (Khattak et al. 2023b) (Li et al. 2024b), we divide the training and testing datasets into base and novel classes. Our teacher model, pre-trained via the PromptSRC approach (Khattak et al. 2023b), guides the training of student model using an unlabeled set. Post-distillation, we assess the students' performance on both class types against the test set, serving as a measure of methodological generalization within the dataset. As shown in Table 1, we compare the performance of our proposed ComKD-CLIP with recent state-of-the-art methods including CLIP (Radford et al. 2021), CoOp (Zhou et al. 2022b), CoCoOp (Zhou et al. 2022a), MaPLe (Khattak et al. 2023a), PromptSRC (Khattak et al. 2023b), PromptKD (Li et al. 2024b) on 11 recognition datasets. In comparison with these state-of-the-art works, ComKD-CLIP shows highly competitive results in all 11 datasets.

Cross-dataset Evaluation

Similar to PromptKD (Li et al. 2024b), our teacher model undergoes pre-training on ImageNet (Deng et al. 2009). Subsequently, we utilize the training set of unlabeled target datasets to train the student model. Their performance is then evaluated on the test set post-training, with no data-specific fine-tuning applied. We compare our cross-dataset performance with previous methods in Table 2, our proposed ComKD-CLIP outperforms some state-of-the-art methods on 8 out of 10 datasets, achieving an average improvement of 0.74% over previous methods.

Domain Generalization Experiments

We train a source model on the ImageNet (Deng et al. 2009) dataset and subsequently evaluate its robustness across various out-of-distribution datasets to assess performance under domain shifts. This method helps us explore the model's adaptability to different and unexpected environments, thereby identifying its strengths and potential vulnerabilities in practical applications. We summarize the results

Table 1: We compare base-to-novel generalization capabilities with current state-of-the-art methods. Our **ComKD-CLIP** framework demonstrates exceptional generalization across 11 recognition datasets, employing the **ViT-B/16 image encoder** from the CLIP model. The symbol Δ represents the performance improvements relative to the previously established state-of-the-art, **PromptKD**.

(a) Average over 11 datasets

ViT-B/16	Base	Novel	HM
CLIP	69.34	74.22	71.70
CoOp	82.69	63.22	71.66
CoCoOp	80.47	71.69	75.83
MaPLe	82.28	75.14	78.55
PromptSRC	84.26	76.10	79.97
PromptKD	86.96	80.73	83.73
ComKD-CLIP	87.37	80.59	83.84
Δ	**+0.41**	-0.14	**+0.11**

(b) Caltech101

ViT-B/16	Base	Novel	HM
CLIP	96.84	94.00	95.40
CoOp	98.00	89.81	93.73
CoCoOp	97.96	93.81	95.84
MaPLe	97.74	94.36	96.02
PromptSRC	98.10	94.03	96.02
PromptKD	98.91	96.65	97.77
ComKD-CLIP	99.23	96.40	97.79
Δ	**+0.32**	-0.25	**+0.02**

(c) OxfordPets

ViT-B/16	Base	Novel	HM
CLIP	91.17	97.26	94.12
CoOp	93.67	95.29	94.47
CoCoOp	95.20	97.69	96.43
MaPLe	95.43	97.76	96.58
PromptSRC	95.33	97.30	96.30
PromptKD	96.30	98.01	97.15
ComKD-CLIP	96.76	98.10	97.43
Δ	**+0.46**	**+0.09**	**+0.28**

(d) Flowers102

ViT-B/16	Base	Novel	HM
CLIP	72.08	77.80	74.83
CoOp	97.60	59.67	74.06
CoCoOp	94.87	71.75	81.71
MaPLe	95.92	72.46	82.56
PromptSRC	98.07	76.50	85.95
PromptKD	99.42	82.62	90.24
ComKD-CLIP	99.53	82.62	90.29
Δ	**+0.11**	+0.00	**+0.05**

(e) Food101

ViT-B/16	Base	Novel	HM
CLIP	90.10	91.22	90.66
CoOp	88.33	82.26	85.19
CoCoOp	90.70	91.29	90.99
MaPLe	90.71	92.05	91.38
PromptSRC	90.67	91.53	91.10
PromptKD	92.43	93.68	93.05
ComKD-CLIP	92.78	93.91	93.34
Δ	**+0.35**	**+0.23**	**+0.29**

(f) FGVCAircraft

ViT-B/16	Base	Novel	HM
CLIP	27.19	36.29	31.09
CoOp	40.44	22.30	28.75
CoCoOp	33.41	23.71	27.74
MaPLe	37.44	35.61	36.50
PromptSRC	42.73	37.87	40.15
PromptKD	49.12	41.81	45.17
ComKD-CLIP	51.80	43.37	47.21
Δ	**+2.68**	**+1.56**	**+2.04**

(g) SUN397

ViT-B/16	Base	Novel	HM
CLIP	69.36	75.35	72.23
CoOp	80.60	65.89	72.51
CoCoOp	79.74	76.86	78.27
MaPLe	80.82	78.70	79.75
PromptSRC	82.67	78.47	80.52
PromptKD	83.69	81.54	82.60
ComKD-CLIP	84.19	81.70	82.93
Δ	**+0.50**	**+0.16**	**+0.33**

(h) DTD

ViT-B/16	Base	Novel	HM
CLIP	53.24	59.90	56.37
CoOp	79.44	41.18	54.24
CoCoOp	77.01	56.00	64.85
MaPLe	80.36	59.18	68.16
PromptSRC	83.37	62.97	71.75
PromptKD	85.84	71.37	77.94
ComKD-CLIP	86.46	70.89	77.90
Δ	**+0.62**	-0.48	-0.04

(i) EuroSAT

ViT-B/16	Base	Novel	HM
CLIP	56.48	64.05	60.03
CoOp	92.19	54.74	68.69
CoCoOp	87.49	60.04	71.21
MaPLe	94.07	73.23	82.35
PromptSRC	92.90	73.90	82.32
PromptKD	97.54	82.08	89.14
ComKD-CLIP	98.17	81.79	89.23
Δ	**+0.63**	-0.29	**+0.09**

(j) UCF101

ViT-B/16	Base	Novel	HM
CLIP	70.53	77.50	73.85
CoOp	84.69	56.05	67.46
CoCoOp	82.33	73.45	77.64
MaPLe	83.00	78.66	80.77
PromptSRC	87.10	78.80	82.74
PromptKD	89.71	82.27	86.10
ComKD-CLIP	90.28	81.45	85.64
Δ	**+0.57**	-0.82	-0.46

(k) StanfordCars

ViT-B/16	Base	Novel	HM
CLIP	63.37	74.89	68.65
CoOp	78.12	60.40	68.13
CoCoOp	70.49	73.59	72.01
MaPLe	72.94	74.00	73.47
PromptSRC	78.27	74.97	76.58
PromptKD	82.80	83.37	83.13
ComKD-CLIP	85.33	85.19	85.26
Δ	**+2.53**	**+1.82**	**+2.13**

(l) ImageNet

ViT-B/16	Base	Novel	HM
CLIP	72.43	68.14	70.22
CoOp	76.47	67.88	71.92
CoCoOp	75.98	70.43	73.10
MaPLe	76.66	70.54	73.47
PromptSRC	77.60	70.73	74.01
PromptKD	80.83	74.66	77.62
ComKD-CLIP	76.53	71.02	73.67
Δ	-4.30	-3.64	-3.95

of ComKD-CLIP and compare with previous methods on out-of-distribution datasets in Table 3. The results indicate that ComKD-CLIP outperforms some state-of-the-art methods on the source datasets ImageNetV2, ImageNet-Sketch, and ImageNetA, with higher average performance as well. This demonstrates that ComKD-CLIP has stronger generalization capabilities on datasets with domain shifts.

Ablation Study

The effectiveness of IFAlign & EduAttention

To begin with, we try to explore the effectiveness of IFAlign module and EduAttention module. Specifically, we remove the IFAlign module and EduAttention module from ComKD-CLIP respectively and test the corresponding experimental results. As shown in Table 4, removing IFAlign module will greatly reduce the performance of ComKD-CLIP. Similarly, removing the EduAttention module also decrease the model's performance. It is worth noting that when both the IFAlign module and the EduAttention module are removed, ComKD-CLIP has the worst performance. These experimental results fully prove that IFAlign module and EduAttention module can effectively prompt the student model to absorb the teacher model's knowledge on how to extract image features and integrate text-image feature.

Subsequently, we try to use different alignment methods in IFAlign to find the best alignment strategy. We respectively align image features by mean ($\mathcal{L}_{\text{align\_mean}}$), vari-

Table 2: The performance of **ComKD-CLIP** with some state-of-the-art mothods in a cross-dataset benchmark. Utilizing our pipeline, we conduct unsupervised Aligned distillation with unlabeled domain data in a transductive setting. The source model is pretrained on ImageNet. "ZSL" indicates that the evaluation is conducted in a Zero-Shot Learning setting. **ComKD-CLIP** outperforms on 8 out of 10 datasets.

ZSL	ViT-B/16	Target Dataset										
		Caltech 101	Oxford Pets	Flowers 102	Food101	FGVC Aircraft	SUN397	DTD	Euro SAT	UCF101	Stanford Cars	Avg.
In-ductive	CoOp	93.70	89.14	68.71	85.30	18.47	64.15	41.92	46.39	66.55	64.51	63.88
	CoCoOp	94.43	90.14	71.88	86.06	22.94	67.36	45.73	45.37	68.21	65.32	65.74
	MaPLe	93.53	90.49	72.23	86.20	24.74	67.01	46.49	48.06	68.69	65.57	66.30
	PromptSRC	93.60	90.25	70.25	86.15	23.90	67.10	46.87	45.50	68.75	65.70	65.81
Trans-ductive	PromptKD	93.61	**91.59**	**75.33**	88.84	26.24	68.57	55.08	63.74	76.39	73.93	71.33
	ComKD-CLIP	**94.56**	91.36	75.07	**89.21**	**27.54**	**69.82**	**57.09**	**64.27**	**77.16**	**74.59**	**72.07**
	Δ	**+0.95**	-0.23	-0.26	**+0.37**	**+1.30**	**+1.25**	**+2.01**	**+0.53**	**+0.77**	**+0.66**	**+0.74**

Table 3: The performance of ComKD-CLIP within domain generalization contexts. The method is trained on ImageNet and subsequently evaluated on datasets with domain shifts.

ViT-B/16	Target Dataset				
	-V2	-S	-A	-R	Avg.
CLIP	60.83	46.15	47.77	73.96	57.18
CoOp	64.20	47.99	49.71	75.21	59.28
CoCoOp	64.07	48.75	50.63	76.18	59.91
MapLe	64.07	49.15	50.90	76.98	60.27
PromptSRC	64.35	49.55	50.90	77.80	60.65
PromptKD	69.77	58.72	70.36	**87.01**	71.47
ComKD-CLIP	**70.95**	**59.92**	**74.15**	86.08	**72.78**
Δ	**+1.18**	**+1.2**	**+3.79**	-0.93	**+1.31**

Table 4: Ablation study of IFAlign module and EduAttention module on SUN397.

Methods	Base	Novel	HM
ComKD-CLIP w/o IFAlign	83.89	81.68	82.77
ComKD-CLIP w/o EduAttention	83.83	**81.70**	82.75
ComKD-CLIP w/o IFAlign w/o EduAttention	83.69	81.54	82.60
ComKD-CLIP (Full)	**84.19**	**81.70**	**82.93**

ance ($\mathcal{L}_{\text{align\_var}}$), and mean add variance ($\mathcal{L}_{\text{align}}$). The experimental results are shown in Table 5. It can be clearly seen that ComKD-CLIP has the best performance when we use the mean add variance ($\mathcal{L}_{\text{align}}$) to align image features in IFAlign. This mainly because that mean add variance can jointly promote the alignment from the perspectives of center position and discreteness, thereby helping the student model to absorb the teacher model's knowledge on how to extract image features.

Table 5: Ablation study of alignment strategy used in IFAlign on SUN397.

Methods	Base	Novel	HM
$\mathcal{L}_{\text{align\_var}}$	83.88	81.21	82.52
$\mathcal{L}_{\text{align\_mean}}$	83.84	**81.93**	82.87
$\mathcal{L}_{\text{align}}$	**84.19**	81.70	**82.93**

Knowledge Refinement Method

Finally, we try to explore the best strategy to refine the knowledge distilled by IFAlign module and EduAttention module. Specifically, we use KL divergence, L1 and MSE methods to refine the distilled knowledge. The corresponding performance of ComKD-CLIP is shown in Table 6. It can be seen that KL divergence can better refine the distilled knowledge and make ComKD-CLIP perform best. This may be because KL divergence can make the logits of the student model better approximate the logits of the teacher model, thus more effectively refining the knowledge distilled by the student model.

Table 6: Comparison between different distillation methods in SUN397.

KD Loss	Base	Novel	HM
L1	66.16	58.84	62.29
MSE	67.79	61.93	64.73
KL	**84.19**	**81.70**	**82.93**

Conclusion

In this study, we present ComKD-CLIP, an advanced knowledge distillation framework designed to comprehensively distill the knowledge from a large teacher CLIP model to a smaller student model. This process maintains the comparable performance while substantially reducing the model's parameter. ComKD-CLIP innovatively employs IFAlign and EduAttention to effectively distill the intricate knowledge embedded within the teacher CLIP model during the text-image feature fusion process. Futhermore, ComKD-CLIP can refine the distilled knowledge by leveraging the feature fusion results of the teacher model, ensuring that the smaller student model can accurately absorb the knowledge from teacher model. Extensive experiments across 11 datasets have unequivocally demonstrated the superior performance of ComKD-CLIP. The proposed method significantly bolsters the capabilities of smaller CLIP models in resource-constrained environments, thereby broadening the practical utility of the CLIP technology.

Acknowledgments

This research was supported by the Natural Science Basic Research Program of Shaanxi (2024JC-YBMS-468) and the China National Key R&D Program (2022YFC2808003).

References

Ali, M.; and Khan, S. 2023. Clip-decoder: Zeroshot multilabel classification using multimodal clip aligned representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4675–4679.

Bossard, L.; Guillaumin, M.; and Van Gool, L. 2014. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, 446–461. Springer.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3606–3613.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dong, X.; Bao, J.; Zheng, Y.; Zhang, T.; Chen, D.; Yang, H.; Zeng, M.; Zhang, W.; Yuan, L.; Chen, D.; et al. 2023. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10995–11005.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2004. 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*, 178–178. IEEE.

Gao, Y.; Liu, J.; Xu, Z.; Zhang, J.; Li, K.; Ji, R.; and Shen, C. 2022. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *Advances in neural information processing systems*, 35: 35959–35970.

Helber, P.; Bischke, B.; Dengel, A.; and Borth, D. 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7): 2217–2226.

Hendrycks, D.; Basart, S.; Mu, N.; Kadavath, S.; Wang, F.; Dorundo, E.; Desai, R.; Zhu, T.; Parajuli, S.; Guo, M.; et al. 2021a. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8340–8349.

Hendrycks, D.; Zhao, K.; Basart, S.; Steinhardt, J.; and Song, D. 2021b. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15262–15271.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Jia, Z.; Sun, S.; Liu, G.; and Liu, B. 2024. MSSD: multi-scale self-distillation for object detection. *Visual Intelligence*, 2(1): 8.

Khattak, M. U.; Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2023a. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19113–19122.

Khattak, M. U.; Wasim, S. T.; Naseer, M.; Khan, S.; Yang, M.-H.; and Khan, F. S. 2023b. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15190–15200.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, 554–561.

Laroudie, C.; Bursuc, A.; Ha, M. L.; and Franchi, G. 2023. Improving clip robustness with knowledge distillation and self-training. *arXiv preprint arXiv:2309.10361*.

Li, Y.; Fan, H.; Hu, R.; Feichtenhofer, C.; and He, K. 2023a. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23390–23400.

Li, Y.; Pang, G.; Suo, W.; Jing, C.; Xi, Y.; Liu, L.; Chen, H.; Liang, G.; and Wang, P. 2024a. CoLeCLIP: Open-Domain Continual Learning via Joint Task Prompt and Vocabulary Learning. *arXiv preprint arXiv:2403.10245*.

Li, Z.; Li, X.; Fu, X.; Zhang, X.; Wang, W.; Chen, S.; and Yang, J. 2024b. Promptkd: Unsupervised prompt distillation for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26617–26626.

Li, Z.; Li, X.; Yang, L.; Zhao, B.; Song, R.; Luo, L.; Li, J.; and Yang, J. 2023b. Curriculum temperature for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1504–1512.

Li, Z.; Ye, J.; Song, M.; Huang, Y.; and Pan, Z. 2021. Online knowledge distillation for efficient pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11740–11750.

Lin, H.; Bai, H.; Liu, Z.; Hou, L.; Sun, M.; Song, L.; Wei, Y.; and Sun, Z. 2024. Mope-clip: Structured pruning for efficient vision-language models with module-wise pruning error metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27370–27380.

Liu, Y.; Chen, K.; Liu, C.; Qin, Z.; Luo, Z.; and Wang, J. 2019. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2604–2613.

Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; and Vedaldi, A. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

Mu, N.; Kirillov, A.; Wagner, D.; and Xie, S. 2022. Slip: Self-supervision meets language-image pre-training. In *European conference on computer vision*, 529–544. Springer.

Nilsback, M.-E.; and Zisserman, A. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, 722–729. IEEE.

Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; and Jawahar, C. 2012. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, 3498–3505. IEEE.

Pei, R.; Liu, J.; Li, W.; Shao, B.; Xu, S.; Dai, P.; Lu, J.; and Yan, Y. 2023. Clipping: Distilling clip-based models with a student base for video-language retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18983–18992.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Recht, B.; Roelofs, R.; Schmidt, L.; and Shankar, V. 2019. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, 5389–5400. PMLR.

Singh, A.; Hu, R.; Goswami, V.; Couairon, G.; Galuba, W.; Rohrbach, M.; and Kiela, D. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15638–15650.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Wang, H.; Ge, S.; Lipton, Z.; and Xing, E. P. 2019. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32.

Wang, J.; Chen, Y.; Zheng, Z.; Li, X.; Cheng, M.-M.; and Hou, Q. 2024. CrossKD: Cross-head knowledge distillation for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16520–16530.

Wang, Z.; Codella, N.; Chen, Y.-C.; Zhou, L.; Yang, J.; Dai, X.; Xiao, B.; You, H.; Chang, S.-F.; and Yuan, L. 2022. Clip-td: Clip targeted distillation for vision-language tasks. *arXiv preprint arXiv:2201.05729*.

Wei, L.; Xie, L.; Zhou, W.; Li, H.; and Tian, Q. 2022. Mvp: Multimodality-guided visual pre-training. In *European conference on computer vision*, 337–353. Springer.

Wu, K.; Peng, H.; Zhou, Z.; Xiao, B.; Liu, M.; Yuan, L.; Xuan, H.; Valenzuela, M.; Chen, X. S.; Wang, X.; et al. 2023. Tinyclip: Clip distillation via affinity mimicking and weight inheritance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21970–21980.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, 3485–3492. IEEE.

Yang, C.; An, Z.; Huang, L.; Bi, J.; Yu, X.; Yang, H.; Diao, B.; and Xu, Y. 2024. CLIP-KD: An Empirical Study of CLIP Model Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15952–15962.

Yang, C.; Zhou, H.; An, Z.; Jiang, X.; Xu, Y.; and Zhang, Q. 2022. Cross-image relational knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12319–12328.

Yang, Y.; Huang, W.; Wei, Y.; Peng, H.; Jiang, X.; Jiang, H.; Wei, F.; Wang, Y.; Hu, H.; Qiu, L.; et al. 2023. Attentive mask clip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2771–2781.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhu, X.; Zhang, R.; He, B.; Zhou, A.; Wang, D.; Zhao, B.; and Gao, P. 2023. Not all features matter: Enhancing few-shot clip with adaptive prior refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2605–2615.