CLIP-KD: An Empirical Study of CLIP Model Distillation

Chuanguang Yang[1,2]   Zhulin An[1*]   Libo Huang[1]   Junyu Bi[1,2]   Xinqiang Yu[1,2]

Han Yang[1,2]   Boyu Diao[1]   Yongjun Xu[1*]

[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

[2]University of Chinese Academy of Sciences, Beijing, China

{yangchuanguang, anzhulin, huanglibo, bijunyu, yuxinqiang21s}@ict.ac.cn

{yanghan22s, diaoboyu2012, xyj}@ict.ac.cn

Abstract

Contrastive Language-Image Pre-training (CLIP) has become a promising language-supervised visual pre-training framework. This paper aims to distill small CLIP models supervised by a large teacher CLIP model. We propose several distillation strategies, including relation, feature, gradient and contrastive paradigms, to examine the effectiveness of CLIP-Knowledge Distillation (KD). We show that a simple feature mimicry with Mean Squared Error loss works surprisingly well. Moreover, interactive contrastive learning across teacher and student encoders is also effective in performance improvement. We explain that the success of CLIP-KD can be attributed to maximizing the feature similarity between teacher and student. The unified method is applied to distill several student models trained on CC3M+12M. CLIP-KD improves student CLIP models consistently over zero-shot ImageNet classification and cross-modal retrieval benchmarks. When using ViT-L/14 pretrained on Laion-400M as the teacher, CLIP-KD achieves 57.5% and 55.4% zero-shot top-1 ImageNet accuracy over ViT-B/16 and ResNet-50, surpassing the original CLIP without KD by 20.5% and 20.1% margins, respectively. Our code is released on https://github.com/winycg/CLIP-KD.

1. Introduction

Language-supervised image pre-training has attracted widespread attention for visual representation learning. As a representative work, CLIP (Contrastive Language-Image Pre-training) [36] applies contrastive learning to (image, text) pairs. It guides the model to predict the correct (image, text) pair among the candidate image and text samples. Pre-trained CLIP models show excellent versatility in zero-shot multimodal and unimodal visual tasks.

Some recent works improve CLIP using an extra visual self-supervision task [23, 33] or mask images [24, 55]. The pre-trained CLIP model is also introduced as a remarkable teacher to guide downstream visual pre-training [21, 35, 46]. However, few previous works explore improving the valuable small CLIP models in resource-constrained applications. This paper introduces CLIP-Knowledge Distillation (KD), which aims to enhance a small student CLIP model supervised by a pre-trained large teacher CLIP model. The state-of-the-art TinyCLIP [48] also investigates CLIP distillation. A critical core of TinyCLIP is weight inheritance, which transfers part weights from the well-trained teacher model to a smaller student model. However, this mechanism needs the teacher and student models to have the same architecture-style, *e.g.*, ViT-B/32 [8] to ViT-61M/32 and ResNet-101 [12] to ResNet-30M, limiting the scope of practical applications. This paper provides a comprehensive study on distilling small CLIP models from relation, feature, gradient, and contrastive paradigms. Our CLIP-KD does not rely on architectural-cue and can generalize to any teacher-student architecture pair.

Given the teacher and student CLIP models, we design distillation strategies from the view of mimicry and interaction. For mimicry learning, we guide the student to align the corresponding knowledge generated from the teacher, which is a basic framework in KD [3, 15, 38]. The core question is how to construct meaningful knowledge. Under CLIP, we build contrastive image-to-text relationships, (image, text) features and gradients for mimicry. For interactive learning, we combine the teacher and student for joint contrastive learning, letting the student learn from the teacher implicitly. For example, the student is regarded as an anchor to contrast the teacher embeddings. We also aggregate the student and teacher features for CLIP training.

We train CLIP models over Conceptual [4, 40] datasets and evaluate pre-trained models over zero-shot ImageNet [5] classification and cross-modal retrieval on MSCOCO [29] and Flickr [57]. All proposed distilla-

---

* Corresponding author

tion methods improve the student CLIP models with various margins. Surprisingly, a simple feature mimicry with Mean Squared Error loss could achieve the best performance. Moreover, interactive contrastive learning fulfills the second-best performance. We find that the distillation performance is in line with how much the feature similarity between teacher and student is maximized. This explains why various KD methods have different performance. The unified method is used to distill a series of student networks with different architectures and achieves consistent improvements. For example, when trained on CC3M+12M, CLIP-KD improves MobileViT-S from 32.6% to 36.0% on zero-shot ImageNet accuracy, reducing the gap with the teacher ViT-B/16's 37.0%. When using ViT-L/14 pretrained on Laion-400M [39] as the teacher, CLIP-KD increases ViT-B/16 trained on CC3M+12M with a 20.5% zero-shot ImageNet accuracy gain compared to the baseline.

The main contributions are summarized as follows:

- We propose several distillation strategies, including relation, feature, gradient and contrastive paradigms, to examine the effectiveness of CLIP-KD. A simple feature mimicry loss works surprisingly well. Interactive contrastive learning also achieves good performance.
- We explain that a good CLIP distillation method could maximize the feature similarity between teacher and student models. Intuitively, if the student's features perfectly align with the teacher's features, their performance gap could disappear.
- We provide comprehensive guidelines for CLIP-KD. Compared to state-of-the-art TinyCLIP [48], our CLIP-KD does not rely on architecture-cue and achieves better performance on both the same- and different-architecture styles of the teacher-student models.

2. Related Works

Language-Supervised Learning. Some previous multi-modal works explore visual representations supervised by language. A critical problem is how to create meaningful interaction between visual and linguistic. CLIP [36] is a representative approach using contrastive learning over image-text pairs. ALIGN [18] utilizes larger-scale contrastive pairs with noisy text supervision. Contrastive multi-modal learning [9, 58–60] has popularized exploring cross-modal correlation. Beyond the contrastive paradigm, generative approaches [1, 6, 44] have been examined for visual-linguistic learning. Our method focuses on CLIP distillation that improves the performance of the small CLIP models.

CLIP Variants. Some recent works attempt to improve CLIP with better performance and efficiency. SLIP [33] combines CLIP and visual self-supervised learning as a multi-task framework. MaskCLIP [7] introduces mask self-distillation to train an image EMA encoder for CLIP. DeCLIP [23] performs data-efficient pre-training through multi-dimension supervision signals. Beyond auxiliary supervision, FLIP [24] and A-CLIP [55] conduct image masking over the input to accelerate CLIP training and achieve a better trade-off between performance and efficiency. In contrast, our paper focuses on CLIP compression using KD instead of a new CLIP method.

Multi-Modal Knowledge Distillation. Knowledge Distillation (KD) [15] has been applied to a broad range of tasks, such as visual recognition [17, 25, 26, 50, 52], language model compression [19], and multi-modal representation learning [10, 27]. DistillVLM [10] aligns hidden attention distributions and feature maps between teacher and student. This simple yet effective idea has been applied to many multi-modal KD works [22, 28, 45]. Recently, Tiny-CLIP [48] also aims for CLIP distillation and achieves satisfactory performance via affinity mimicking and weight inheritance. However, the weight inheritance mechanism requires the same architecture-style between teacher and student models. By contrast, CLIP-KD could adapt any architecture pair without considering architectural correlation.

3. Methodology

3.1. A Brief Review of CLIP

CLIP (Contrastive Language-Image Pre-Training). Given a set of (image, text) pairs denoted as $\mathcal{D} = \{(I_k, T_k)\}_{k=1}^{|\mathcal{D}|}$, CLIP performs an image-text alignment task to push the paired image-text close and unpaired ones apart in the feature embedding space. The CLIP framework includes a visual encoder f_i and a text encoder f_t to transform the image I_k and the text T_k into feature embeddings v_k and s_k respectively, $i.e.$ $v_k = f_i(I_k)$, $s_k = f_t(T_k)$. Here, all embeddings are post-processed by l_2 normalization. CLIP adopts InfoNCE-based [34] contrastive loss to maximize the similarity between v_k and s_k against other negative samples. Given the image embedding v_k as the anchor, the image-to-text contrastive loss is formulated as:

$$\mathcal{L}_{I \to T} = -\log \frac{\exp(v_k \cdot s_k / \tau)}{\sum_{b=1}^{|\mathcal{B}|} \exp(v_k \cdot s_b / \tau)}. \quad (1)$$

CLIP conducts a symmetric image-text alignment contrastive loss. Given the text embedding s_k as the anchor, the text-to-image contrastive loss is formulated as:

$$\mathcal{L}_{T \to I} = -\log \frac{\exp(s_k \cdot v_k / \tau)}{\sum_{b=1}^{|\mathcal{B}|} \exp(s_k \cdot v_b / \tau)}. \quad (2)$$

Here, $\cdot$ denotes the dot product to measure the similarity, τ is a learnable temperature to scale the distribution. In practice, the negative samples are retrieved from the mini-batch $\mathcal{B}$. The total loss of CLIP is formulated as:

$$\mathcal{L}_{CLIP} = \frac{1}{2}(\mathcal{L}_{I \to T} + \mathcal{L}_{T \to I}). \quad (3)$$

(a) Contrastive Relational Distillation (CRD).

(b) Feature Distillation (FD).

(c) Masked Feature Distillation (MFD).

(d) Gradient Distillation (GD).

(e) Interactive Contrastive Learning (ICL).

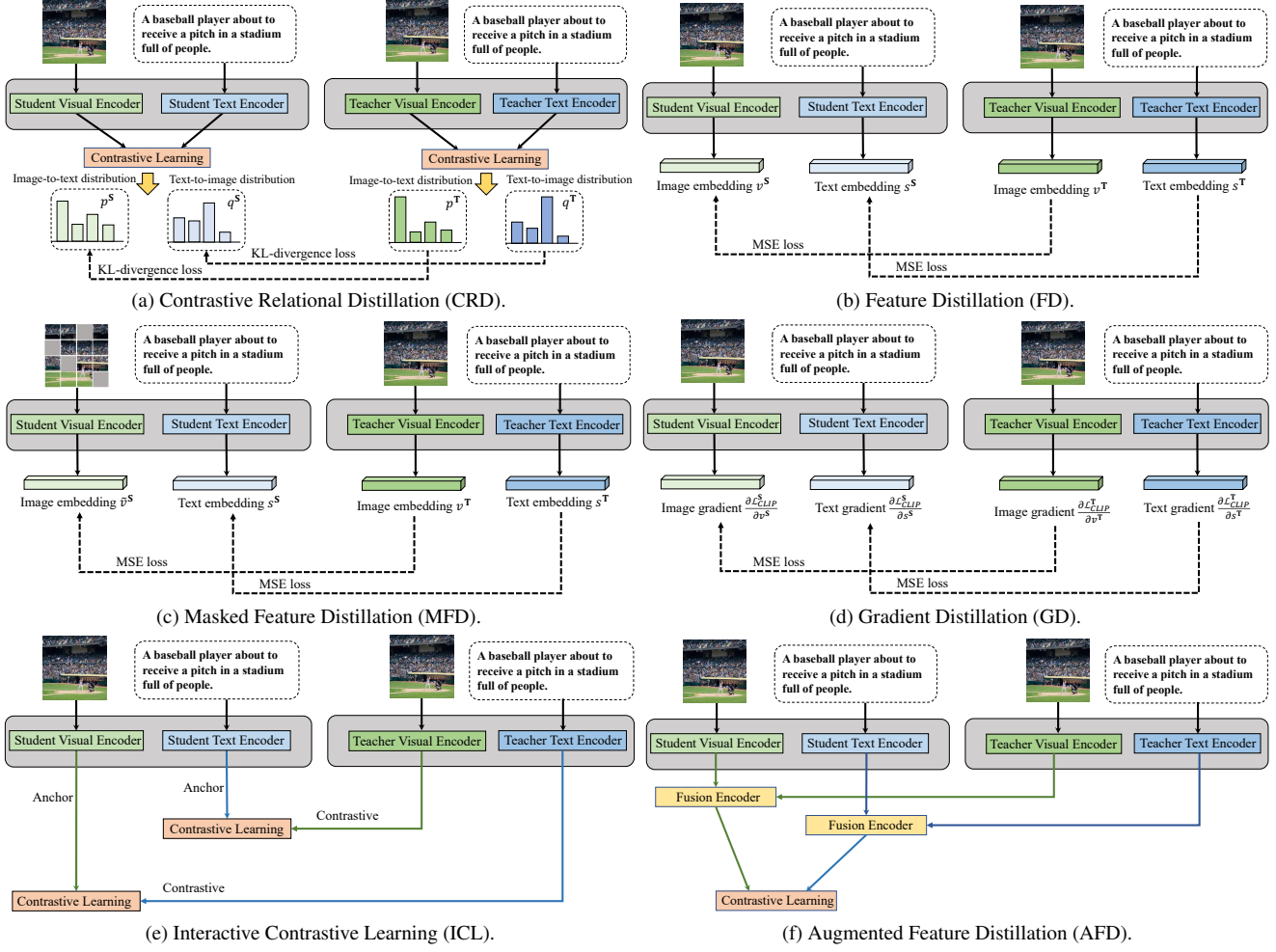(f) Augmented Feature Distillation (AFD).

Figure 1. Illustration of various CLIP knowledge distillation approaches proposed in this paper.

3.2. CLIP Knowledge Distillation

In this section, we propose several CLIP distillation methods and illustrate the overview of details in Fig. 1.

3.2.1 Contrastive Relational Distillation

The core idea of CLIP is to maximize the similarity between the paired image-text embeddings over the contrastive similarity distribution. Therefore, the straightforward knowledge type is output-oriented contrastive distribution for Contrastive Relational Distillation (CRD). This idea is also used by some previous works for image classification [11, 51, 54], object detection [56] and semantic segmentation [53]. The contrastive distribution captures the structured relationships among feature embeddings. A good teacher often constructs a well-structured feature space. CRD makes the student mimic better structured semantic relations from the teacher, further improving the quality of feature representations.

Given a mini-batch $\mathcal{B} = \{(I_k, T_k)\}_{k=1}^{|\mathcal{B}|}$, the generated (visual, text) embeddings from teacher and student are $\{(v_k^{\mathbf{T}}, s_k^{\mathbf{T}})\}_{k=1}^{|\mathcal{B}|}$ and $\{(v_k^{\mathbf{S}}, s_k^{\mathbf{S}})\}_{k=1}^{|\mathcal{B}|}$, respectively. Given the k-th image embedding v_k as an anchor, the teacher and student image-to-text contrastive distributions $p_k^{\mathbf{T}} \in \mathbb{R}^{|\mathcal{B}|}$ and $p_k^{\mathbf{S}} \in \mathbb{R}^{|\mathcal{B}|}$ are formulated as:

$$p_k^{\mathbf{T}}[j] = \frac{\exp(v_k^{\mathbf{T}} \cdot s_j^{\mathbf{T}} / \tau^{\mathbf{T}})}{\sum_{b=1}^{|\mathcal{B}|} \exp(v_k^{\mathbf{T}} \cdot s_b^{\mathbf{T}} / \tau^{\mathbf{T}})}, \tag{4}$$

$$p_k^{\mathbf{S}}[j] = \frac{\exp(v_k^{\mathbf{S}} \cdot s_j^{\mathbf{S}} / \tau^{\mathbf{S}})}{\sum_{b=1}^{|\mathcal{B}|} \exp(v_k^{\mathbf{S}} \cdot s_b^{\mathbf{S}} / \tau^{\mathbf{S}})}. \tag{5}$$

Here, $j \in [1, 2, \cdots, |\mathcal{B}|]$ denotes the index of the contrastive distribution. Symmetrically, given the text embedding s_k as an anchor, the teacher and student text-to-image contrastive distributions $q_k^{\mathbf{T}} \in \mathbb{R}^{|\mathcal{B}|}$ and $q_k^{\mathbf{S}} \in \mathbb{R}^{|\mathcal{B}|}$ are formulated as:

$$q_k^{\mathbf{T}}[j] = \frac{\exp(s_k^{\mathbf{T}} \cdot v_j^{\mathbf{T}} / \tau^{\mathbf{T}})}{\sum_{b=1}^{|\mathcal{B}|} \exp(s_k^{\mathbf{T}} \cdot v_b^{\mathbf{T}} / \tau^{\mathbf{T}})}, \tag{6}$$

$$q_k^{\mathbf{S}}[j] = \frac{\exp(s_k^{\mathbf{S}} \cdot v_j^{\mathbf{S}}/\tau^{\mathbf{S}})}{\sum_{b=1}^{|\mathcal{B}|} \exp(s_k^{\mathbf{S}} \cdot v_b^{\mathbf{S}}/\tau^{\mathbf{S}})}. \quad (7)$$

We align the contrastive distributions between teacher and student via KL-divergence loss. For image-to-text and text-to-image, the distillation losses are formulated as Eq.(8) and Eq.(9):

$$\mathcal{L}_{CRD\_I \to T} = \frac{1}{|\mathcal{B}|} \sum_{k=1}^{|\mathcal{B}|} \sum_{j=1}^{|\mathcal{B}|} p_k^{\mathbf{T}}[j] \log \frac{p_k^{\mathbf{T}}[j]}{p_k^{\mathbf{S}}[j]}, \quad (8)$$

$$\mathcal{L}_{CRD\_T \to I} = \frac{1}{|\mathcal{B}|} \sum_{k=1}^{|\mathcal{B}|} \sum_{j=1}^{|\mathcal{B}|} q_k^{\mathbf{T}}[j] \log \frac{q_k^{\mathbf{T}}[j]}{q_k^{\mathbf{S}}[j]}. \quad (9)$$

The total CRD loss for CLIP distillation is summarized as:

$$\mathcal{L}_{CRD} = \mathcal{L}_{CRD\_I \to T} + \mathcal{L}_{CRD\_T \to I}. \quad (10)$$

3.2.2 Feature Distillation

A simple yet effective way is to align feature embeddings between teacher and student to reduce the knowledge gap directly. Intuitively, if the student's features perfectly align with the teacher's features, their performance gap could disappear. We guide the student to mimic the teacher's visual and text embeddings via Mean Squared Error (MSE) loss:

$$\mathcal{L}_{FD} = \frac{1}{|\mathcal{B}|} \sum_{k=1}^{|\mathcal{B}|} (\|v_k^{\mathbf{T}} - v_k^{\mathbf{S}}\|_2^2 + \|s_k^{\mathbf{T}} - s_k^{\mathbf{S}}\|_2^2). \quad (11)$$

Here, when the embedding sizes between teacher and student are different, we apply a linear projection head to student embeddings to match the dimension.

3.2.3 Masked Feature Distillation

The core idea of masked image modeling [2, 13, 49] is to recover the masked regions using contextual information modeling by a vision transformer. In the scenario of distillation, the teacher is a good supervisor that could provide valuable information to help the student recover the visual semantics given the masked image as input. Like FD, we utilize MSE loss to align the student's and teacher's visual and text embeddings. The difference is that Masked Feature Distillation (MFD) uses masked images as the input to a student. The patch masking algorithm is followed from MAE [13]. The total loss of MFD is formulated as:

$$\mathcal{L}_{MFD} = \frac{1}{|\mathcal{B}|} \sum_{k=1}^{|\mathcal{B}|} (\|v_k^{\mathbf{T}} - \tilde{v}_k^{\mathbf{S}}\|_2^2 + \|s_k^{\mathbf{T}} - s_k^{\mathbf{S}}\|_2^2), \quad (12)$$

where $\tilde{v}_k^{\mathbf{S}}$ is the visual embedding based on the masked input image.

3.2.4 Gradient Distillation

The gradient information often shows how the model responds to changes according to inputs. We propose to force the gradient consistency between teacher and student using the derivative w.r.t the visual and text embeddings. By this means, the student could better understand how the output should change according to the input. This helps the student behave more similarly to the teacher.

We align the gradient information w.r.t each visual and text embedding between teacher and student via MSE loss:

$$\mathcal{L}_{GD} = \frac{1}{|\mathcal{B}|} \sum_{k=1}^{|\mathcal{B}|} (\left\| \frac{\partial \mathcal{L}_{CLIP}^{\mathbf{T}}}{\partial v_k^{\mathbf{T}}} - \frac{\partial \mathcal{L}_{CLIP}^{\mathbf{S}}}{\partial v_k^{\mathbf{S}}} \right\|_2^2$$
$$+ \left\| \frac{\partial \mathcal{L}_{CLIP}^{\mathbf{T}}}{\partial s_k^{\mathbf{T}}} - \frac{\partial \mathcal{L}_{CLIP}^{\mathbf{S}}}{\partial s_k^{\mathbf{S}}} \right\|_2^2). \quad (13)$$

Derivations in Eq.(13) are shown in Appendix Section 1.

3.2.5 Interactive Contrastive Learning

To facilitate the interaction between teacher and student, we propose *Interactive Contrastive Learning* (ICL) across the student and teacher feature encoders. It regards the student as an anchor to contrast the teacher's embeddings. Given the student image embedding $v_k^{\mathbf{S}}$, the contrastive text embeddings denoted as $\{s_b^{\mathbf{T}}\}_{b=1}^{|\mathcal{B}|}$ are from the teacher text encoder instead of the student text encoder. The image-to-text ICL loss is formulated as:

$$\mathcal{L}_{ICL\_I \to T} = -\log \frac{\exp(v_k^{\mathbf{S}} \cdot s_k^{\mathbf{T}}/\tau)}{\sum_{b=1}^{|\mathcal{B}|} \exp(v_k^{\mathbf{S}} \cdot s_b^{\mathbf{T}}/\tau)}. \quad (14)$$

Symmetrically, given the student text embedding $s_k^{\mathbf{S}}$, the contrastive image embeddings denoted as $\{v_b^{\mathbf{T}}\}_{b=1}^{|\mathcal{B}|}$ are from the teacher visual encoder. The text-to-image ICL loss is formulated as:

$$\mathcal{L}_{ICL\_T \to I} = -\log \frac{\exp(s_k^{\mathbf{S}} \cdot v_k^{\mathbf{T}}/\tau)}{\sum_{b=1}^{|\mathcal{B}|} \exp(s_k^{\mathbf{S}} \cdot v_b^{\mathbf{T}}/\tau)}. \quad (15)$$

The total loss of ICL is summarized as:

$$\mathcal{L}_{ICL} = \frac{1}{2}(\mathcal{L}_{ICL\_I \to T} + \mathcal{L}_{ICL\_T \to I}). \quad (16)$$

We demonstrate that minimizing $\mathcal{L}_{ICL}$ is connected to maximizing the lower bound of mutual information between teacher and student networks. The mutual information measures the uncertainty reduction in contrastive feature embeddings from the teacher network when the anchor embedding from the student network is known. By maximizing the lower bound of mutual information, the student network learns more common knowledge from the teacher network, leading to better feature representations. *The theoretical proof is shown in Appendix Section 2.*

Table 1. **Comparison of CLIP distillation losses trained from CC3M+12M on zero-shot ImageNet-related classification and cross-modal retrieval on CC3M Val, MSCOCO and Flickr.** The numbers in **bold** denote the best results for individual methods (the third block) and unified methods (the fourth block), respectively. The 'T' and 'S' tags represent the teacher and student roles, respectively.

Method	IN Acc	INV2 Acc	IN-R Acc	IN-S Acc	CC3M Val I2T	CC3M Val T2I	MSCOCO I2T	MSCOCO T2I	Flickr I2T	Flickr T2I
T: ViT-B/16	37.0	32.1	48.4	26.0	40.2	39.5	25.0	24.7	54.6	56.6
S: ViT-T/16	30.6	25.6	35.7	17.4	33.3	33.3	20.7	20.3	46.4	47.7
+CRD	31.9	27.6	38.8	19.6	35.3	34.9	21.4	20.7	48.8	49.9
+FD	**34.2**	**29.5**	**42.7**	**21.4**	37.1	**36.9**	22.5	**22.2**	**51.1**	**51.3**
+MFD	34.1	**29.5**	42.3	21.2	**37.4**	**36.9**	**22.9**	22.1	50.9	51.1
+GD	31.5	27.0	37.9	19.1	34.5	34.0	21.3	20.9	47.5	48.3
+ICL	33.1	28.2	40.6	20.8	36.1	35.8	21.8	21.7	50.5	50.4
+AFD	31.4	26.9	37.8	18.6	34.6	34.7	20.9	20.5	47.3	48.7
+FD+ICL	34.6	30.0	43.2	**22.0**	37.9	37.6	23.0	22.5	51.7	51.9
+FD+ICL+CRD	**34.9**	**30.1**	43.5	21.9	**38.2**	**37.9**	23.1	**22.6**	52.3	**52.4**
+FD+ICL+CRD+GD	34.8	29.9	42.8	**22.0**	38.1	37.7	**23.3**	22.5	**52.4**	52.3
+FD+ICL+CRD+AFD	34.8	**30.1**	**43.6**	21.6	**38.2**	37.7	23.0	22.5	52.2	**52.4**

3.2.6 Augmented Feature Distillation

To help the interaction between teacher and student, we propose to augment the student embeddings using the teacher embeddings by a fusion encoder. We hope the teacher can guide the student to optimize a meaningful visual-text embedding space. We introduce a visual fusion encoder ϕ_i and a text fusion encoder ϕ_t to aggregate the student and teacher embeddings:

$$v_k^{\mathbf{A}} = \phi_i(v_k^{\mathbf{S}}||v_k^{\mathbf{T}}), \quad s_k^{\mathbf{A}} = \phi_t(s_k^{\mathbf{S}}||s_k^{\mathbf{T}}). \quad (17)$$

Here, $||$ is the concatenation operator, and the fusion encoder is a simple linear projection layer. The augmented feature embeddings $(v_k^{\mathbf{A}}, s_k^{\mathbf{A}})$ are applied to the general CLIP contrastive loss as Eq.(3).

3.2.7 Overall Loss of CLIP Distillation.

We summarize the original CLIP task and distillation losses to jointly train a student model:

$$\mathcal{L}_{CLIP\_KD} = \mathcal{L}_{CLIP} + \lambda \mathcal{L}_{KD}. \quad (18)$$

Here, $\mathcal{L}_{KD} \in \{\mathcal{L}_{CRD}, \mathcal{L}_{FD}, \mathcal{L}_{MFD}, \mathcal{L}_{GD}, \mathcal{L}_{ICL}, \mathcal{L}_{AFD}\}$ represents a distillation loss. λ is a distillation loss weight to scale the magnitude. Multiple distillation losses can be selectively utilized together.

4. Experiments

4.1. Experimental Setup

Dataset. We use Conceptual Captions 3M (CC3M) [40] and Conceptual 12M (CC12M) [4] for vision-and-language pre-training. We follow the consistent evaluation protocol

Table 2. Configuration of paired visual and text encoders.

Visual encoder			Text encoder: Transformer [42]			
Model	Type	Params	Layer	Width	Head	Params
ViT-L/14 [8]		304.0M	12	768	12	85.1M
ViT-B/16 [8]		86.2M	12	512	8	37.8M
ViT-T/16 [8]	ViT	5.6M				
MobileViT-S [32]		5.3M	12	384	6	21.3M
Swin-T [30]		27.9M				
ResNet-101 [12],		56.3M	12	512	8	37.8M
ResNet-50 [12],		38.3M				
ResNet-18 [12],	CNN	11.4M				
MobileNetV3 [16]		2.0M	12	384	6	21.3M
EfficientNet-B0 [41]		4.7M				

with CLIP-related works [24, 47]. The CC3M validation set, including 13K image-text pairs, is used for cross-modal retrieval evaluation. For zero-shot classification, we utilize the ImageNet (IN) [5] validation set and its several variants, such as ImageNet-V2 (IN-V2) [37], ImageNet-Rendition (IN-R) [14] and ImageNet-Sketch (IN-S) [43] for evaluation. For zero-shot cross-modal image/text retrieval, we adopt MSCOCO [29] and Flickr [57] for evaluation.

Evaluation metrics. Following the standard setting, we employ Recall@K (R@K) to measure the retrieval performance in K nearest neighbours. By default, we use top-1 accuracy (Acc) for image classification and R@1 for Image-to-Text (I2T) and Text-to-Image (T2I) retrieval.

Configuration of visual and text encoders. As shown in Table 2, We show the configuration of visual and text encoders, followed by open_clip[1] codebase.

Training details. We adopt an AdamW optimizer [31] with an initial learning rate of 0.001 and a weight decay of 0.1. A cosine learning rate schedule is applied with a linear warm-up for 10K iterations in 32 epochs. Experiments are

---

[1]https://github.com/mlfoundations/open_clip

Table 3. Distillation performance trained from CC3M+12M for cross-modal retrieval on CC3M, MSCOCO and Flickr validation set.

Method	CC3M		MSCOCO		Flickr		Method	CC3M		MSCOCO		Flickr	
	I2T	T2I	I2T	T2I	I2T	T2I		I2T	T2I	I2T	T2I	I2T	T2I
T: ViT-B/16	40.2	39.5	25.0	24.7	54.6	56.6	T: ResNet-101	41.4	40.5	25.2	25.7	57.0	55.5
S: MobileViT-S	36.0	35.6	22.3	22.9	50.1	53.0	S: MobileViT-S	36.0	35.6	22.3	22.9	50.1	53.0
+CLIP-KD	**39.4**	**38.6**	**26.1**	**24.9**	**55.0**	**56.2**	+CLIP-KD	**39.9**	**38.6**	**26.0**	**25.3**	**57.6**	**56.1**
S: Swin-T	39.8	39.2	24.7	25.3	53.4	54.4	S: Swin-T	39.8	39.2	24.7	25.3	53.4	54.4
+CLIP-KD	**43.7**	**42.5**	**28.5**	**28.6**	**62.2**	**60.9**	+CLIP-KD	**44.2**	**43.0**	**27.8**	**28.9**	**60.8**	**61.5**
S: MobileNetV3	28.1	27.5	15.3	15.0	36.9	38.0	S: MobileNetV3	28.1	27.5	15.3	15.0	36.9	38.0
+CLIP-KD	**30.1**	**28.6**	**17.9**	**16.0**	**42.4**	**42.3**	+CLIP-KD	**30.2**	**29.4**	**17.2**	**16.6**	**40.2**	**42.2**
S: EfficientNet-B0	35.4	34.9	21.7	21.1	48.3	50.1	S: EfficientNet-B0	35.4	34.9	21.7	21.1	48.3	50.1
+CLIP-KD	**39.0**	**38.0**	**26.0**	**23.9**	**55.5**	**54.2**	+CLIP-KD	**37.4**	**36.8**	**24.7**	**24.6**	**55.8**	**56.2**
S: ResNet-18	31.1	30.4	19.2	18.6	41.0	43.3	S: ResNet-18	31.1	30.4	19.2	18.6	41.0	43.3
+CLIP-KD	**34.2**	**33.0**	**21.3**	**19.8**	**47.8**	**47.1**	+CLIP-KD	**34.7**	**33.7**	**21.0**	**20.9**	**48.8**	**48.4**

run over 8 NVIDIA A800 GPUs. The batch size is 1024, where each GPU holds 128 samples. For the weight of each distillation loss, we set $\lambda_{CRD} = 1$, $\lambda_{FD} = \lambda_{MFD} = 2000$, $\lambda_{GD} = 10^8$ and $\lambda_{ICL} = 1$. The learnable temperature τ is initialized from 0.07. Other training settings are followed from the original CLIP [36]. *The detailed hyper-parameter experiments are shown in Appendix Section 3.*

4.2. Ablation Study of Distillation Losses

In this section, we examine the effectiveness of various CLIP distillation approaches. As shown in Table 1, we conduct a comprehensive comparison on zero-shot ImageNet-related classification and cross-modal retrieval. Any individual distillation loss could boost the student performance over the baseline. Feature Distillation (FD) with a simple MSE mimicry loss achieves the best distillation performance among them. It improves the student by 3.6% top-1 accuracy on ImageNet, 3.7%, 1.9% and 4.2% R@1 values on CC3M, MSCOCO and Flickr, respectively. We further evaluate MFD by applying image patch masking into FD. MFD shows similar performance with FD, therefore we do not introduce this technique for CLIP-KD. Beyond MFD, ICL and CRD become the second- and third-best approaches for overall zero-shot performance. GD and AFD lead to relatively moderate performance gains compared to the baseline.

We further combine loss terms to investigate the unified distillation approach. The combination of FD+ICL outperforms the single FD or ICL, indicating that FD and ICL are complementary. We further apply CRD to FD+ICL, and the performance is improved continually. Moreover, we find adding GD or AFD to FD+ICL+CRD may not lead to performance gains. In summary, the combination FD+CRD+ICL performs best in 6 out of 10 cases. By default, we utilize this unified method for distilling various CLIP models in this paper.

4.3. Distilling CLIP Models

Given the pretrained teacher CLIP model, we distill several lightweight student CLIP models with various architectures. The results are evaluated on zero-shot retrieval and ImageNet classification. *We also report linear evaluation experiments on MS-COCO object detection and instance segmentation in Appendix Section 3.*

4.3.1 Cross-Modal Retrieval on CC3M

Table 3 reports distillation performance supervised by ViT-B/16 and ResNet-101 as teachers. The proposed KD approach improves student performance over various network architectures consistently. Supervised by ViT-B/16 for image→text retrieval, KD leads to 2.0%∼3.9% R@1 gains on MobileViT-S, Swin-T, MobileNetV3, EfficientNet-B0 and ResNet-18, respectively. For text→image retrieval, KD results in 1.1%∼3.4% R@1 gains on these networks. Supervised by ResNet-101, KD boosts the baseline by 2.0%∼4.4% R@1 for image→text retrieval, and 1.9%∼3.8% R@1 for text→image retrieval, over these five student networks, respectively. The results demonstrate the effectiveness of CLIP-KD over a series of networks. Moreover, the architectural difference between ViT and CNN does not affect CLIP-KD's performance. This is because our CLIP-KD only considers the final output embeddings for distillation instead of information from hidden layers.

4.3.2 Zero-Shot Cross-Modal Retrieval

As shown in Table 3, we further transfer student CLIP models to zero-shot cross-modal retrieval on MSCOCO and Flickr. Supervised by the teacher ViT-B/16, KD outperforms the baseline by 2.1%∼4.3% MSCOCO R@1 margins for image→text retrieval, and 1.0%∼3.3% MSCOCO R@1 margins for text→image retrieval on various networks. On Flickr, R@1 gains are 4.9%∼8.8% for

Table 4. Distillation performance of zero-shot ImageNet and its variants on top-1 classification accuracy (%) trained on CC3M+12M.

Method	IN-1K	INV2	IN-R	IN-S	Method	IN-1K	INV2	IN-R	IN-S
T: ViT-B/16	37.0	32.1	48.4	26.0	T: ResNet-101	36.8	31.9	49.2	26.7
S: MobileViT-S	32.6	27.6	39.5	20.1	S: MobileViT-S	32.6	27.6	39.5	20.1
+CLIP-KD	**36.0**	**31.1**	**44.5**	**23.5**	+CLIP-KD	**35.0**	**30.1**	**43.7**	**22.7**
S: Swin-T	36.4	31.1	45.9	24.4	S: Swin-T	36.4	31.1	45.9	24.4
+CLIP-KD	**40.2**	**34.9**	**51.4**	**28.2**	+CLIP-KD	**39.5**	**34.2**	**51.9**	**28.1**
S: MobileNetV3	25.1	20.7	29.2	13.4	S:MobileNetV3	25.1	20.7	29.2	13.4
+CLIP-KD	**27.0**	**23.0**	**30.6**	**14.1**	+CLIP-KD	**26.2**	**22.2**	**29.3**	**13.7**
S: EfficientNet-B0	32.6	27.8	40.9	20.7	S:EfficientNet-B0	32.6	27.8	40.9	20.7
+CLIP-KD	**35.4**	**30.6**	**44.7**	**23.7**	+CLIP-KD	**34.6**	**29.4**	**44.4**	**23.1**
S: ResNet-18	28.6	24.0	35.3	18.1	S:ResNet-18	28.6	24.0	35.3	18.1
+CLIP-KD	**31.4**	**26.9**	**39.2**	**20.0**	+CLIP-KD	**30.9**	**25.9**	**38.0**	**19.5**

image→text retrieval, and 3.2%∼6.5% for text→image retrieval. Supervised by the teacher ResNet-101, KD leads to 1.8%∼3.7% R@1 improvements for image→text retrieval, and 1.6%∼3.6% R@1 improvements for text→image retrieval on MSCOCO. On Flickr, KD results in 3.3%∼7.8% R@1 gains for image→text retrieval, and 3.1%∼7.1% for text→image retrieval. The results reveal the transfer ability to zero-shot cross-modal retrieval using CLIP-KD.

4.3.3 Zero-Shot ImageNet-Related Classification

In Table 4, we transfer the student CLIP models to zero-shot ImageNet classification for visual recognition and ImageNet-variants for robustness evaluation. For ImageNet classification supervised by ViT-B/16, KD improves 3.4%, 3.8%, 1.9%, 2.8% and 2.8% top-1 accuracy gains over MobileViT-S, Swin-T, MobileNetV3, EfficientNet-B0 and ResNet-18, respectively. Supervised by ResNet-101, KD achieves 2.4%, 3.1%, 1.1%, 2.0% and 2.3% top-1 accuracy improvements over five networks, respectively. The results show that CLIP-KD can help downstream visual recognition effectively. Extensive experiments over ImageNet variants indicate that CLIP-KD can lead to clear accuracy gains over baseline.

After distillation, Swin-T even outperforms the teacher models. There are two reasons to explain this phenomenon. On the one hand, Swin-T is a powerful model, and the performance gaps with teacher models are small. On the other hand, CLIP-KD transfers meaningful knowledge from teacher models to Swin-T, improving its performance and surpassing teacher models.

4.3.4 Transferred from Laion-400M

Cross-dataset comparison. In Table 5, we use the teachers pretrained from Laion-400M [39] to distill student CLIP models trained on CC3M+12M. We find that the teacher

ViT-B/16 pre-trained on Laion-400M significantly outperforms its counterpart pre-trained on CC3M+12M to distill a student ViT-T/16. It shows a 7.7% ImageNet accuracy gain and an average cross-modal retrieval improvement of 6.8%. The results demonstrate the CLIP-KD can effectively transfer knowledge from a large-scale dataset to improve CLIP models trained on a small-scale dataset. The advantage helps the model learn knowledge from a large-scale dataset without training too many data samples.

Impact of teacher models with different sizes. In Table 5, we use ViT-L/14 or ViT-B/16 as two teachers to investigate the impact of teacher sizes on CLIP-KD. Both of two teachers enhance the student ViT-T/16 over baseline with substantial margins. However, it is counter-intuitive that the more capable ViT-L/14 underperforms the weaker ViT-B/16 for distillation. One possible reason is that a large teacher and a small student may exist capacity gaps, making the student difficult to align with the teacher. This may become an open issue for future research.

Comparison with TinyCLIP. CLIP-KD achieves better performance than state-of-the-art TinyCLIP [48] by 1.8% ImageNet accuracy and 1.7% cross-modal retrieval gains on average. Moreover, we do not provide the results of ResNet-50 for TinyCLIP, because TinyCLIP only supports the teacher and student with the same architecture-style. The results show that CLIP-KD is a more preferable method than TinyCLIP in performance and practicability.

4.4. Analysis

In this section, we conduct thorough analyses and ablation experiments to investigate CLIP-KD. Unless otherwise specified, the teacher and student visual encoders are ViT-B/16 and ViT-T/16, respectively.

Training curve of CLIP-KD As shown in Fig. 2, we illustrate some statistics and analyses of CLIP-KD during the training procedure:

(1) **Training loss analysis.** Fig. 2a shows training curves

Table 5. **Distillation performance of zero-shot ImageNet and cross-modal retrieval trained on CC3M+12M.** The teachers are pretrained on Laion-400M before distillation. '(from T_x)' indicates that the student is distilled from the teacher T_x.

Method	IN-1K Acc	MSCOCO I2T	MSCOCO T2I	Flickr I2T	Flickr T2I
T_1: ViT-L/14	72.8	42.7	40.9	80.5	79.5
T_2: ViT-B/16	67.1	39.5	36.5	76.9	75.5
S: ViT-T/16	30.6	20.7	20.3	46.4	47.7
+TinyCLIP (from T_1)	39.3	26.4	24.1	57.6	57.4
+TinyCLIP (from T_2)	40.8	26.8	24.7	58.6	58.5
+CLIP-KD (from T_1)	40.9	27.2	25.5	59.7	59.7
+CLIP-KD (from T_2)	**42.6**	**28.1**	**26.0**	**60.4**	**59.9**
S: ViT-B/16	37.0	25.0	24.7	54.6	56.6
+TinyCLIP (from T_1)	55.4	35.9	33.6	73.2	72.8
+CLIP-KD (from T_1)	**57.5**	**37.6**	**35.6**	**75.3**	**74.5**
S: ResNet-50	35.3	23.5	24.7	55.1	55.0
+CLIP-KD (from T_2)	**55.4**	**36.3**	**33.4**	**73.0**	**72.2**

(a) Training loss.

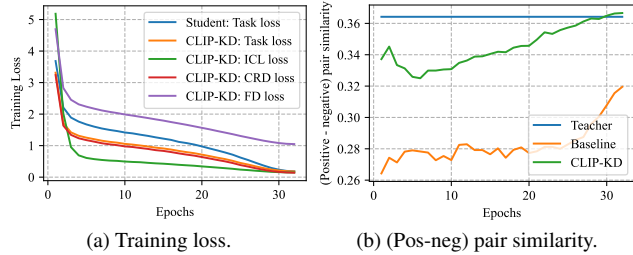(b) (Pos-neg) pair similarity.

Figure 2. Training curves trained on CC3M+12M for CLIP-KD.

of various loss terms. As the training continues, all loss values decrease and then converge until the end of training. CLIP-KD has lower task loss than that of the baseline during the training since it is supervised by a pretrained CLIP teacher. The task loss is often larger than the ICL loss, because the teacher provides converged contrastive embeddings to the student in ICL, helping the student optimize feature space readily.

(2) **Sample similarity analysis.** Fig. 2b shows the similarity curve of positive minus negative pairs, which represents the relative distance between positive and negative pairs. Contrastive learning expects positive pairs to have higher similarities while negative pairs have lower similarities. Both the baseline and CLIP-KD increase (positive-negative) pair similarity during the training stage, indicating a discriminative embedding space is gradually learned. CLIP-KD has higher similarity values than the baseline, manifesting that it guides the student to learn more discriminative features, further benefiting downstream tasks.

Interpreting why various KD methods have different performance. As shown in Figure 3, we analyze various KD methods in different performance from the view of cosine and CKA [20] similarities between student and teacher features after distillation. We find student accuracy is in line

(a) Similarity statistics of image features.

(b) Similarity statistics of text features.

Figure 3. **Similarity statistics between teacher and student features after distillation trained on CC3M+12M.** $v_k^{\mathbf{T}}$ and $v_k^{\mathbf{S}}$ denote the teacher and student image features, respectively. $s_k^{\mathbf{T}}$ and $s_k^{\mathbf{S}}$ denote the teacher and student text features, respectively.

with feature similarity. The larger similarity means that the student learns more similar teacher features, reducing the performance gap with the teacher. The simple FD performs the best because it forces the student to increase the similarity with teacher features directly.

However, FD does not consider informative contrastive image-text relations. ICL is proposed to promote contrastive distillation and increase mutual information between teacher and student, resulting in high similarity. By contrast, CRD, GD, and AFD are relatively weaker in enhancing similarity with the teacher, thus achieving limited gains above baseline. Overall, FD+ICL is capable of feature alignment and contrastive distillation, which is the major source of performance improvement.

5. Conclusion

This paper provides a comprehensive study on CLIP-KD by examining several distillation strategies, including relation, feature, gradient, and contrastive paradigms. Experimental results show that the proposed distillation methods lead to significant improvements on small CLIP models. We hope our study can provide solid CLIP-KD guidelines on practical application and attract more attention to future CLIP compression research.

Acknowledgement

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 2

[2] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 4

[3] Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10925–10934, 2022. 1

[4] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pretraining to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 1, 5

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009. 1, 5

[6] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11162–11173, 2021. 2

[7] Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10995–11005, 2023. 2

[8] 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. *arXiv preprint arXiv:2010.11929*, 2020. 1, 5

[9] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023. 2

[10] 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. 2

[11] Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, and Zicheng Liu. Seed: Self-supervised distillation for visual representation. *ICLR*, 2021. 3

[12] 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. 1, 5

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 4

[14] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 5

[15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1, 2

[16] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019. 5

[17] Libo Huang, Yan Zeng, Chuanguang Yang, Zhulin An, Boyu Diao, and Yongjun Xu. etag: Class-incremental learning via embedding distillation and task-oriented generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. 2

[18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 2

[19] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*, 2019. 2

[20] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR, 2019. 8

[21] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models. *arXiv preprint arXiv:2209.15639*, 2022. 1

[22] Xuanlin Li, Yunhao Fang, Minghua Liu, Zhan Ling, Zhuowen Tu, and Hao Su. Distilling large vision-language model with out-of-distribution generalizability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2492–2503, 2023. 2

[23] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021. 1, 2

[24] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 23390–23400, 2023. 1, 2, 5

[25] Zheng Li, Jingwen Ye, Mingli Song, Ying Huang, and Zhigeng Pan. Online knowledge distillation for efficient pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11740–11750, 2021. 2

[26] Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. Curriculum temperature for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1504–1512, 2023. 2

[27] Zheng Li, Xiang Li, Xinyi Fu, Xin Zhang, Weiqiang Wang, Shuo Chen, and Jian Yang. Promptkd: Unsupervised prompt distillation for vision-language models. *arXiv preprint arXiv:2403.02781*, 2024. 2

[28] Chen Liang, Jiahui Yu, Ming-Hsuan Yang, Matthew Brown, Yin Cui, Tuo Zhao, Boqing Gong, and Tianyi Zhou. Module-wise adaptive distillation for multimodality foundation models. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 5

[30] 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

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[32] Sachin Mehta and Mohammad Rastegari. Mobilevit: lightweight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*, 2021. 5

[33] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pretraining. In *European conference on computer vision*, pages 529–544. Springer, 2022. 1, 2

[34] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[35] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022. 1

[36] 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

[37] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019. 5

[38] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *ICLR*, 2015. 1

[39] 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. 2, 7

[40] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 1, 5

[41] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 5

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[43] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019. 5

[44] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 2

[45] 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. 2

[46] Longhui Wei, Lingxi Xie, Wengang Zhou, Houqiang Li, and Qi Tian. Mvp: Multimodality-guided visual pre-training. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*, pages 337–353. Springer, 2022. 1

[47] Yixuan Wei, Yue Cao, Zheng Zhang, Zhuliang Yao, Zhenda Xie, Han Hu, and Baining Guo. icar: Bridging image classification and image-text alignment for visual recognition. *arXiv preprint arXiv:2204.10760*, 2022. 5

[48] Kan Wu, Houwen Peng, Zhenghong Zhou, Bin Xiao, Mengchen Liu, Lu Yuan, Hong Xuan, Michael Valenzuela, Xi Stephen Chen, Xinggang Wang, et al. Tinyclip: Clip distillation via affinity mimicking and weight inheritance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21970–21980, 2023. 1, 2, 7

[49] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022. 4

[50] Chuanguang Yang, Zhulin An, Linhang Cai, and Yongjun Xu. Hierarchical self-supervised augmented knowledge dis-

tillation. *International Joint Conference on Artificial Intelligence*, pages 1217–1223, 2021. 2

[51] Chuanguang Yang, Zhulin An, Linhang Cai, and Yongjun Xu. Mutual contrastive learning for visual representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3045–3053, 2022. 3

[52] Chuanguang Yang, Zhulin An, Helong Zhou, Linhang Cai, Xiang Zhi, Jiwen Wu, Yongjun Xu, and Qian Zhang. Mixskd: Self-knowledge distillation from mixup for image recognition. In *European Conference on Computer Vision*, pages 534–551. Springer, 2022. 2

[53] Chuanguang Yang, Helong Zhou, Zhulin An, Xue Jiang, Yongjun Xu, and Qian Zhang. Cross-image relational knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12319–12328, 2022. 3

[54] Chuanguang Yang, Zhulin An, Helong Zhou, Fuzhen Zhuang, Yongjun Xu, and Qian Zhang. Online knowledge distillation via mutual contrastive learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10212–10227, 2023. 3

[55] Yifan Yang, Weiquan Huang, Yixuan Wei, Houwen Peng, Xinyang Jiang, Huiqiang Jiang, Fangyun Wei, Yin Wang, Han Hu, Lili Qiu, et al. Attentive mask clip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2771–2781, 2023. 1, 2

[56] Lewei Yao, Renjie Pi, Hang Xu, Wei Zhang, Zhenguo Li, and Tong Zhang. G-detkd: towards general distillation framework for object detectors via contrastive and semantic-guided feature imitation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3591–3600, 2021. 3

[57] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 1, 5

[58] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 2

[59] 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.

[60] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 2