Adapt without Forgetting: Distill Proximity from Dual Teachers in Vision-Language Models

Mengyu Zheng[1,2], Yehui Tang[2], Zhiwei Hao[2,3],
Kai Han[2], Yunhe Wang[2], and Chang Xu[1⋆]

[1] School of Computer Science, Faculty of Engineering, The University of Sydney
[2] Huawei Noah's Ark Lab
[3] School of information and Electronics, Beijing Institute of Technology
mzhe4259@uni.sydney.edu.au, {yehui.tang,kai.han,yunhe.wang}@huawei.com,
haozhw@bit.edu.cn, c.xu@sydney.edu.au

Abstract. Multi-modal models such as CLIP possess remarkable zero-shot transfer capabilities, making them highly effective in continual learning tasks. However, this advantage is severely compromised by catastrophic forgetting, which undermines the valuable zero-shot learning abilities of these models. Existing methods predominantly focus on preserving zero-shot capabilities but often fall short in fully exploiting the rich modal information inherent in multi-modal models. In this paper, we propose a strategy to enhance both the zero-shot transfer ability and adaptability to new data distribution. We introduce a novel graph-based multi-modal proximity distillation approach that preserves the intra- and inter-modal information for visual and textual modalities. This approach is further enhanced with a sample re-weighting mechanism, dynamically adjusting the influence of teachers for each individual sample. Experimental results demonstrate a considerable improvement over existing methodologies, which illustrate the effectiveness of the proposed method in the field of continual learning. Code is available at github.com/myz-ah/AwoForget.

Keywords: Multi-modal model · Continual learning · Zero-shot learning · Graph-based distillation

1 Introduction

Recently, multi-modal pretrained models, particularly exemplified by CLIP [25], have emerged as a cornerstone in generalization capabilities [38]. These models, benefiting from extensive pre-training on large-scale text-image datasets, exhibit two distinct advantages [23, 29, 40]. Firstly, their zero-shot transfer ability enables them to maintain considerable accuracy on previously unseen data, a crucial feature for models expected to adapt to ever-changing real-world scenarios [16, 37]. Secondly, they demonstrate remarkable efficiency in adapting to new downstream tasks with limited training data [35, 43]. The potential of

⋆ Corresponding author

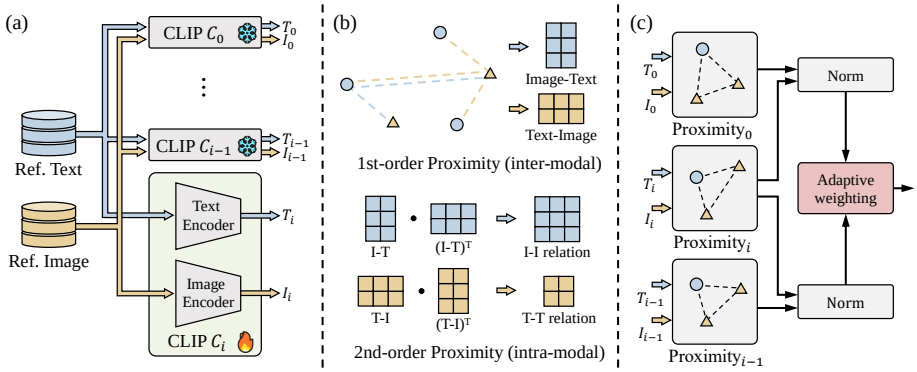

Fig. 1: Overview of the proposed continual learning framework for vision-language models. (a) The dual-teacher distillation with CLIP models C_0 and C_{i-1} using reference data. (b) Cross-modal graph construction for first-order and second-order proximity modeling. (c) The adaptive re-weighting mechanism for balanced knowledge distillation from different teacher.

these multi-modal pretrained models in continual learning (CL) scenarios, where adaptability and long-term learning are essential, appears immense. However, recent research shows that fine-tuning on downstream tasks could damage to the generalization ability of the pretrained models. This decline not only impairs the zero-shot capabilities but also leads to a gradual loss in performance gains on continual learning progress. Moveover, such phenomenon aligns closely with the notorious issue of "catastrophic forgetting" in continual learning [1, 26, 30], posing a significant hurdle in leveraging the full potential of multi-modal models in dynamic and evolving real-world applications.

While replay-based methods and the teacher-student learning paradigm have made significant progress in addressing the decline in generalization capabilities of pretrained models within a continual learning framework [4, 25], they face limitations in balancing knowledge retention and adaptation to new data. Replay-based methods depend on pre-training phase data and labels, struggling to integrate this knowledge with new domain data [20, 27]. Conversely, the teacher-student paradigm, though effective in preserving original capabilities, often fails to exploit the full potential of rich multi-modal information, particularly in adapting to diverse and evolving datasets. Approaches like ZSCL [45] within the teacher-student framework introduce reference data to mitigate forgetting, providing valuable insights but not fully addressing the dynamic complexity of multi-modal information.

In addressing the identified shortcomings of existing continual learning strategies for multi-modal models, our research emphasizes alleviating the conflict between maintaining zero-shot transfer abilities and adapting to new datasets. This challenge is significant as preserving zero-shot capabilities often impedes a model's plasticity to new data. Our study introduces an innovative dual-teacher distillation strategy, which focuses on maintaining the relative proximity between

samples rather than directly modeling individual outputs. Relative relationships are more conducive to preserving a model's inherent capabilities while ensuring its adaptability. To effectively model these proximity, we introduce a sophisticated graph-based approach that captures both intra- and inter-modal interactions. Then a proximity distillation method is conducted with this graph to learn and preserve relationships from teachers. Additionally, we propose a sample re-weighting mechanism designed to dynamically balance the insights from the two teacher models from sample aspect, which allows us to further address the tension between model stability and plasticity. We conclude the framework of the proposed method in Figure 1.

In summary, our approach revolutionizes continual learning for multi-modal models by introducing a graph-based multi-modal proximity distillation coupled with an innovative sample re-weighting mechanism. A key contribution of our work is the novel use of a graph structure to reveal both inter- and intra-modal relative relationships for CLIP models. This methodology not only resolves the inherent conflict between maintaining zero-shot capabilities and adapting to new datasets but also significantly enhances the adaptability and stability of multi-modal models in continue learning learning environments. Experimental results on several benchmarks illustrate the superiority of the proposed method.

2 Related Work

Vision Language Model. Inspired by human multi-modal learning process, pretrained vision-language model(VLMs) models have been proposed and are rapidly gaining significant attention. In comparison to vision models [9,12] , such as ViT [5] , VLMs could be applied in a broader range of downstream tasks [8,47] because of the stunning zero-shot learning capability [44]. In particular, Contrastive Language-Vision Pre-training(CLIP) has a simple model structure but excellent performance [25]. As capturing rich knowledge from massive image-text pairs, CLIP could even performs well on zero-shot classification without fine-tuning. Apparently, it is essential to maintain zero-shot learning capacity in downstream tasks.

Continual Learning. Continual Learning(CL) methods are able to continually learn a long series of tasks, and perform well on all learned tasks without forgetting old ones [17]. How to resisting catastrophic forgetting is the maintain problem in CL. There are three principal kinds of approaches: replay-based, regularization-based [13, 17] and parameter isolation methods [7, 21]. Continual learning based on CLIP model shows impressive performance [32], primarily due to zero-shot learning capabilities. AttriCLIP [33] introduces trainable attribute prompts to mitigate the catastrophic forgetting problem. To retain zero-shot transfer, existing methods apply regularization loss, including distillation loss. LWF-VR [4] distilled generated sentences from the vocabularies to simulate both zero-shot and previous classes. A teacher selection mechanism based on Euclidean distance is proposed to keep zero-shot capability [41]. Moreover, ZSCL [45] introduced reference images and reference sentences, learning KD-

divergence from CLIP in every task. Nonetheless, methods above only directly import extra data without considering about keeping zero-shot and remembering old knowledge in previous tasks jointly.

Knowledge Distillation. Knowledge distillation was proposed for model compression initially [2,19,39], which is also an efficient method for preventing catastrophic forgetting in Continual Learning [4]. Hinton *et al.* [10] presents a teacher-student framework to allow knowledge transformed from pretrained large model to student model. Different from some works only concentrating on transforming instance features, such as final predicted probabilities [4], and features of intermediate layers [28, 42]. Several existing works pay attention to the images relationships [18, 24]. Unlike such distilling relationship approaches, our method is able to maintain first-order proximity and second-order proximity to capture intra- and inter-modal interactions.

3 Preliminary

CLIP on Continual Learning. CLIP [25] is renowned as a multimodal model, distinguished by its exceptional zero-shot transfer ability for various downstream tasks. Unlike traditional computer vision systems, such as ResNet, which categorize images using predefined labels, CLIP integrates a text encoder. This encoder effectively translates labels described in natural language into text features, aligning them with corresponding visual representations. The CLIP model undergoes pre-training on a vast corpus of over 400 million image-text pairs, leveraging a strategy that synergizes these two modalities for improved performance and versatility. Specifically, the training objective is defined as:

$$L_{\mathrm{C}} = -\sum_{i=1}^{N} \log \frac{\exp(\mathrm{sim}(E_{\mathrm{img}}(\boldsymbol{x}_i), E_{\mathrm{txt}}(\boldsymbol{y})/\tau)}{\sum_{j=1}^{N} \exp(\mathrm{sim}(E_{\mathrm{img}}(\boldsymbol{x}_i), E_{\mathrm{txt}}(\boldsymbol{t}_j))/\tau)}, \tag{1}$$

where $\mathrm{sim}(u, v) = u^T v / \|u\| \cdot \|v\|$ is the cosine similarity function, τ is a temperature parameter, $\boldsymbol{y}$ is the text that correctly pairs with image $\boldsymbol{x}$, and $E_{\mathrm{img}}(\boldsymbol{x}_i)$ and $E_{\mathrm{txt}}(\boldsymbol{t}_j)$ are the embeddings from the image and text encoders for the i-th and j-th elements in a batch, respectively. This approach enables the model to effectively bridge the semantic gap between visual and textual data, fostering a robust multimodal understanding.

In the context of continual learning, the CLIP model is sequentially fine-tuned across a series of tasks, each associated with distinct datasets denoted as $\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_N$. A crucial objective in this process is for CLIP to preserve its inherent zero-shot capabilities and simultaneously mitigate catastrophic forgetting. In this paper, the fine-tuned model for each specific task $\mathcal{T}_i$ is represented by C_i, which is iteratively built upon the foundation of the preceding model C_{i-1}. For reference, the original pre-trained CLIP model is designated as C_0.

Knowledge Distillation in Data-Unknown Scenarios. The concept of conducting knowledge distillation in scenarios where the training data of teacher model is unknown, as highlighted in works like wild distillation [3] and the

ZSCL [45] framework, is of notable relevance to the field. This approach gains significance in the context of models like CLIP, characterized by extensive yet undefined pre-training datasets. Here, the student network $\mathcal{N}_S$ is tailored to leverage 'wild' data, often unaligned with the teacher's training regime. The distillation process adapts to this complexity as:

$$L_{\mathrm{KD}} = \sum_{i=0} [L_{\mathrm{CE}} \left(\mathcal{N}_T(\boldsymbol{x}_i, \boldsymbol{t}), \mathcal{N}_S(\boldsymbol{x}_i, \boldsymbol{t})\right)], \tag{2}$$

where $\boldsymbol{x} \sim p_{\mathrm{wild}}(\boldsymbol{x})$ and $\boldsymbol{t} \sim p_{\mathrm{wild}}(\boldsymbol{t})$ denote the diverse, unstructured data samples and L_{CE} represents the cross-entropy loss function. This paradigm, resonating with ZSCL's emphasis on leveraging varied data for continual learning, showcases the necessity of refining student models to function effectively in data-ambiguous environments, underscoring a nuanced approach to knowledge distillation.

4 Method

In this section, we outline our proximity distillation strategy to learn from two teacher models. We construct a graph utilizing multi-modal information and define both first-order and second-order proximity. Then we employ a teacher-student paradigm, focusing on distilling the proximity defined on this graph from both the original CLIP model and a model fine-tuned on previous datasets. To balance the above two teachers and further alleviate the conflict between stability and plasticity, a sample re-weighting mechanism is incorporated to our method.

4.1 Graph-Based Multi-modal Representation

As mentioned above, preserving relationships between samples rather than their individual features, offers a more flexible constraint, allowing the model ample freedom to adapt to varying data distributions while maintaining its inherent capabilities. In this part, we develop a cross-modal graph where each node represents a sample, and the edges reflect the model's depiction of relationships between these samples across different modalities.

We define this graph as $G = (V_I, V_T, E)$, where V_I and V_T respectively represent sets of from image and text modalities. Each vertex is categorized either as an image vertex $v_I \in V_I$ or a text vertex $v_T \in V_T$, with the edges $e \in E$ being undirected to illustrating the mutual interaction that exist between the visual and textual components. This graph is constructed to encapsulate the relationships inherent in multi-modal model. It serves as a foundation for our advanced manifold distillation technique, which we elaborate on in the following sections.

First-Order Proximity. A fundamental component of our methodology is the integration of first-order proximity within the framework of the CLIP model to optimize its performance in continual learning contexts. We utilize a graph $G = (V_T, V_I, E)$ to capture the direct pairwise relationships between the visual and textual modalities. The concept of first-order proximity is essential in

quantifying the immediate similarity between two vertices, thereby facilitating a understanding of these modality interactions.

In graph embedding learning [31], first-order proximity is typically defined by the probability of two vertices sharing similarities. It can be mathematically represented as:

$$\hat{p}_1(v_i, v_j) = \text{sim}(\boldsymbol{u}_i, \boldsymbol{u}_j), \quad \forall (i,j) \in E, \tag{3}$$

where $\boldsymbol{u}_i$ and $\boldsymbol{u}_j$ are the embeddings of the respective vertices, and $\text{sim}(u, v) = \frac{u^T v}{\|u\| \cdot \|v\|}$ is the cosine similarity function as illustrated above. The first-order proximity is defined as zero for any pair of vertices that are not connected by an edge.

It is important to note that within the CLIP model, direct connections only occur between vertices from two distinct modalities. Consequently, the first-order proximity effectively aligns with the classification mechanism of the CLIP model. This is achieved by extracting embeddings for each modality using their respective encoders, such as $\boldsymbol{u}_I = \text{Enc}_I(v_I)$ for image vertices and $\boldsymbol{u}_T = \text{Enc}_T(v_T)$ for textual vertices. In this context, a higher value of $\hat{p}_1(v_i, v_j)$ denotes a stronger correlation between the vertices, indicating a higher likelihood of the image sample v_i being associated with the label represented by text vertex v_j.

Through the lens of first-order proximity, we define the relationships between image and text modalities within our graph-based framework. This approach not only characterizes the classification mechanism of the model but also serves as a flexible regularization. This form of proximity is particularly informative in preserving the capability of a model, making it a valuable asset in our continual learning strategy.

Second-Order Proximity. Building on the first-order proximity that focuses on immediate and inter-modal proximity in our multi-modal framework, we integrate second-order proximity [31] to model a broader and global perspective. This broader view enables us to delve into intra-modal proximity within the CLIP model, going beyond direct image-text connections.

Second-order proximity assesses vertex similarity by examining neighborhood structures. We initially define the proximity between two directly connected vertices v_i and v_j through first-order proximity in Eq. (3). From this foundation, the conditional distribution for a vertex v_i with a neighboring vertex v_n is derived as:

$$p_1(v_n|v_i) = \frac{\text{sim}(\boldsymbol{u}_i, \boldsymbol{u}_n)}{\sum_{(i,k)\in E} \text{sim}(\boldsymbol{u}_i, \boldsymbol{u}_k)}, \quad \forall (i,n) \in E. \tag{4}$$

Then we can model the second-order proximity between two vertices without direct connections by measuring the similarity in their conditional distributions across respective neighborhoods. Formally, we define the second-order proximity between two vertices v_i and v_j as:

$$\hat{p}_2(v_j, v_i) = \text{sim}(p_1(\cdot|v_i), p_1(\cdot|v_j)), \quad \forall (i,j) \notin E. \tag{5}$$

Eq. (5) highlights that second-order proximity identifies similarities between unconnected vertices based on neighborhood likeness. In the CLIP model, connections are established between every image-text vertex pair. Consequently, unconnected vertices typically belong to the same modality, making second-order proximity a measure of intra-modal proximity.

Our approach leverages the multi-modal proximity established by the CLIP model to construct intra-modal connections. Specifically, in Eq. (5), we define the proximity between vertices of the same modality, such as images, utilizing the contextual associations provided by the textual modality. This strategy differs from the standard pairwise similarity often calculated as $\boldsymbol{u}_i \cdot \boldsymbol{u}_j^T$, which typically does not incorporate multi-modal information. Through this, we effectively utilize the unique strengths of the CLIP model in capturing cross-modal interactions to enhance the modeling of same-modality proximity.

Comparison with existing knowledge distillation methods. To elucidate the distinctions between our method and existing knowledge distillation approaches, we primarily focus on the calculation of cosine similarity between image nodes, particularly emphasizing the relationships defined within our framework. Our method diverges from the standard normalization process to streamline the explanation and clarity of the proximity representation. Consequently, we can reformulate Eq. (5) as follows,

$$p_2'(v_j, v_i) = \boldsymbol{u}_i^T \cdot \boldsymbol{U}_T \boldsymbol{U}_T^T \cdot \boldsymbol{u}_j, \quad \forall (i,j) \notin E, \tag{6}$$

where $\boldsymbol{U}_T = [\boldsymbol{u}_{T_0}, \boldsymbol{u}_{T_1}, ..., \boldsymbol{u}_{T_N}]$ donates the embeddings for all text vertices v_T extracted by the corresponding encoder. Eq. (6) demonstrates that our methodology for second-order proximity extends beyond conventional pairwise comparisons, which typically achieved through cosine similarity and denoted as $\boldsymbol{u}_i \cdot \boldsymbol{u}_j^T$. By integrating weighted similarities, where the weights are informed by shared connections with text vertices, we provide a more comprehensive analysis of relationships. This method distinctly differs from and surpasses the conventional $\boldsymbol{u}_i \cdot \boldsymbol{u}_j^T$ approach by incorporating multi-modal information.

In summary, The foundational distinction of our method from existing knowledge distillation techniques lies in its dual-focus on inter-modal and intra-modal relations. Firstly, our approach introduces the concept of first-order proximity to encapsulate inter-modal relations. Secondly, it explores second-order proximity by examining the similarity of neighborhood structures, rather than direct feature similarity, thereby providing a more granular understanding of intra-modal relations. This exploration of neighborhood structures represents a significant departure from existing methods, which primarily focus on direct feature comparisons. In the distillation phase, we utilize a conditional distribution model for second-order proximity, as shown below,

$$p_2(v_j|v_i) = \frac{\exp(p_2(v_i, v_j))}{\sum_{(i,k)\notin E} \exp(p_2(v_i, v_k))}, \quad \forall (i,j) \notin E. \tag{7}$$

This refined approach, grounded in a comprehensive examination of both inter-modal and intra-modal relationships, underscores our method's innovation and its departure from traditional knowledge distillation frameworks.

4.2 Dual Distillation with Samplewise Balance

In the domain of continual learning, maintaining the zero-shot transfer capabilities of the CLIP model while adapting to novel data distributions is of paramount

importance. Zero-shot transfer encapsulates the model's ability to generalize to untrained data, whereas adaptation evaluates its performance post fine-tuning on novel datasets. This dual requirement often leads to a dichotomy, where fine-tuning to optimize for a specific dataset might compromise the model's generalization ability.

To address this, we introduce a dual-teacher distillation framework that capitalizes on the distinct advantages of the pre-trained CLIP model C_0 and its preceding iteration C_{i-1}. This approach is anchored in leveraging unpaired image and text data for distillation in the absence of original pre-training data, inspired by wild distillation [3] and ZSCL [45] methodologies. This approach enables effective knowledge transfer even without access to the original multimodal training dataset.

In our approach, a key objective is for the student model to preserve both first and second order proximities as exhibited by the teacher model, using a set of reference data. This can be achieved through the following objective function:

$$\overline{\ell}_{\mathrm{pd}}(v_k) = L_{\mathrm{CE}}\left(\overline{p}_1(\cdot|v_k), p_1(\cdot|v_k)\right) + L_{\mathrm{CE}}\left(\overline{p}_2(\cdot|v_k), p_2(\cdot|v_k)\right), \tag{8}$$

where $\overline{p}_1(\cdot|v_k)$ and $\overline{p}_2(\cdot|v_k)$ represent the first-order and second-order proximity of the original CLIP model, respectively. In parallel, we utilize $\tilde{p}_1$ and $\tilde{p}_2$ to denote the first-order and second-order proximity within the model C_{i-1}. It is straightforward to distill proximity from two teachers as following objective function,

$$L_{\mathrm{PD}} = \frac{1}{N}\sum_{k=1}^{N}[\overline{\ell}_{\mathrm{pd}}(v_k) + \lambda \cdot \tilde{\ell}_{\mathrm{pd}}(v_k)]. \tag{9}$$

Eq.(9) tries to balance the knowledge from two teachers applying a uniform weighting across different samples, which does not account for the individual learning needs or the diversity of the dataset, potentially limiting the efficacy of the distillation process. To address this, we introduces a dynamic adjustment to the samples, towards a effective knowledge transfer tailored to the specific needs of each sample. Specifically, for a sample v_k, the distance between teacher models C_0 and student C_i is assessed as follows,

$$\overline{d}_p(v_k) = d\left(\overline{p}_1(v_k), p_1(v_k)\right), \tag{10}$$

where $d(\cdot,\cdot)$ denotes the distance metric. The distance $\tilde{d}_p(v_k)$ between model C_{i-1} and C_i is similarly defined. Then the samplewise weights for the k-th sample are normalized accordingly,

$$\overline{w}_k = \frac{\exp(\overline{d}_p(v_k))}{\exp(\overline{d}_p(v_k)) + \exp(\tilde{d}_p(v_k))}. \tag{11}$$

The weight $\tilde{w}_k$ for C_{i-1} model is also similarly defined. This strategy tailors the learning focus for each sample in the reference dataset, determining whether to

align with C_0 for zero-shot capabilities or C_{i-1} for the adaptability. Then the proximity distillation loss can be reformulated as

$$L_{\mathrm{PD}} = \frac{1}{N}\sum_{k=1}^{N}[\overline{w}_k \cdot \bar{\ell}_{\mathrm{pd}}(v_k) + (1-\overline{w}_k)\cdot \tilde{\ell}_{\mathrm{pd}}(v_k)]. \tag{12}$$

This adaptive sample-wise methodology not only preserves zero-shot abilities acquired during pre-training but also allows for effective adaptation to new data distribution, crucial for performance in dynamic learning environments. The final objective function encapsulates our approach:

$$L = L_{\mathrm{PD}} + L_{\mathrm{CE}}. \tag{13}$$

The second term L_{CE} serves as the original objective function for fine-tuning the CLIP model with a new dataset. Paired with our introduced first term L_{PD}, this combined methodology aims to better maintain the zero-shot learning capabilities of CLIP while it undergoes fine-tuning on novel datasets. We have chosen to apply equal weighting to both loss components because experimental results demonstrates low sensitivity to variations in weighting.

5 Experiment

5.1 Experimental setup

Continual learning setting. We mainly evaluate the proposed method on two continental learning settings. They are multi-domain task incremental learning (MTIL) and class incremental learning (CIL).

MTIL [46] is a benchmark of cross-domain version of task incremental learning, where different tasks are sourced from different datasets. This benchmark demands models to learn knowledge from different domains to achieve promising performance. Specifically, MTIL benchmark consists of 11 datasets, which are Aircraft, Caltech101, CIFAR100, DTD, EuroSAT, Flowers, Food, MNIST, OxfordPet, StanfordCars, and SUN397. Each dataset serves as an individual task for continental learning. We follow the setting in [46] and utilize two orders of tasks for evaluation. The first one (Order I) follows the alphabetical order, while the second one (Order II) follows a random order.

CIL is a essential setting for continual learning, requiring the model to be updated incrementally with only new class instances [22]. We evaluate our method on CIFAR100 [14] and TinyImageNet [15] datasets by following the same evaluation setting in [6]. Specifically, with a t-step setting, classes in CIFAR100 are split into t groups evenly and each group serves as a task. We set $t = \{10, 20, 50\}$. Besides, 100 classes are learned in TinyImageNet at a base step and the remaining 100 classes are split into t groups and we set $t = \{5, 10, 20\}$ for TinyImageNet.

Metric. Three metrics are adopted in our experiments, which are "Transfer", "Avg." and "Last". If there are N continual learning tasks in total, and the model

Table 1: Comparison results on the MTIL benchmark under task order setting I and II. Δ represents the performance gap between a CLIP model after continual learning and the original CLIP model (Zero-shot).

	Order I						Order II					
Method	**Transfer**	Δ	**Avg.**	Δ	**Last**	Δ	**Transfer**	Δ	**Avg.**	Δ	**Last**	Δ
Zero-shot	69.4	-	65.3	-	65.3	-	65.4	-	65.3	-	65.3	-
Continual FT	44.6	-24.8	55.9	-9.4	77.3	+12.0	46.6	-18.8	56.2	-9.1	67.4	+2.1
LwF [17]	56.9	-12.5	64.7	-0.6	74.6	+9.0	53.2	-12.2	62.2	-5.2	71.9	+6.6
iCaRL [27]	50.4	-19.0	65.7	+0.4	80.1	+14.8	50.9	-14.5	56.9	-8.4	71.6	+6.3
LwF-VR [4]	57.2	-12.2	65.1	-0.2	76.6	+11.3	50.9	-14.5	56.9	-8.4	71.6	+6.3
WiSE-FT [34]	52.3	-17.1	60.7	-4.6	77.7	+12.4	51.0	-14.4	61.5	-5.9	72.2	+6.9
ZSCL [46]	68.1	-1.3	75.4	+10.1	83.6	+18.3	64.2	-1.2	74.5	+9.2	83.4	+18.1
Ours	**69.8**	**+0.4**	**76.9**	**+11.6**	**85.1**	**+19.8**	**65.4**	**+0.0**	**75.9**	**+10.6**	**85.4**	**+20.1**

achieves an accuracy of $\mathcal{A}(i|j)$ on task i after training on task j, then we can formulate these metrics as:

$$\begin{aligned} \text{Transfer} &= \frac{1}{N-1}\sum_{j=2}^{N}\frac{1}{j-1}\sum_{i=1}^{j}\mathcal{A}(i|j), \\ \text{Avg.} = \frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N}\mathcal{A}(i|j), &\quad \text{Last} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{A}(i|N). \end{aligned} \tag{14}$$

The Avg. and the Last metrics are adopted across all our experiments. On the MTIL benchmark, we also adopt the Transfer metric to evaluate zero-shot transfer ability of a trained model.

Model. We employ a pretrained CLIP [25] model with a ViT-B/16 image encoder in our experiments. Input resolution and patch size are 224 and 16.

Baseline. Since the CLIP model can adapt to new data distributions without retraining, we adopt zero-shot transfer as one of the baseline [32]. Direct fine-tuning the CLIP model on each task is also invited, which is condidered as an upper bound. Besides, we compare the proposed method with not only conventional continual learning methods, but also recent proposed method WiSE-FT [34], LwF-VR [4] and ZSCL [46], which are specifically designed for CLIP.

5.2 Results on MTIL

We compare the proposed method with baselines on the MTIL benchmark. The CLIP model is trained on each task for 1000 iterations with a batch size of 64. The learning rate is configured at 3e-5 for the initial task and 1e-5 for all subsequent tasks. Adam is adopted as the optimizer.

Table 1 showcases the results under both task order settings I and II. As the results show, the proposed method outperforms all other baselines under every metric. Notably, when task order settings I is adopted, the Transfer performance of our proposed method is 69.8%, which is 0.4% higher than even the original CLIP model. Meanwhile all existing methods demonstrate declined

Table 2: Performance of trained model on each task in the MTIL benchmark under task order setting I.

Method	Aircraft	Caltech101	CIFAR100	DTD	EuroSAT	Flowers	Food	MNIST	OxfordPet	Cars	SUN397
Zero-shot	24.3	88.4	68.2	44.6	54.9	71.0	88.5	59.4	89.0	64.7	65.2
Fine-tuning	62.0	95.1	89.6	79.5	98.9	97.5	92.7	99.6	94.7	89.6	81.8
Transfer											
Continual FT		67.1	46.0	32.1	35.6	35.0	57.7	44.1	60.8	20.5	46.6
LwF [17]		74.5	56.9	39.1	51.1	52.6	72.8	60.6	75.1	30.3	55.9
iCaRL [27]		56.6	44.6	32.7	39.3	46.6	68.0	46.0	77.4	31.9	60.5
LwF-VR [4]		77.1	61.0	40.5	45.3	54.4	74.6	47.9	76.7	36.3	58.6
WiSE-FT [34]		73.5	55.6	35.6	41.5	47.0	68.3	53.9	69.3	26.8	51.9
ZSCL [46]		86.0	67.4	45.4	50.4	69.1	87.6	61.8	86.8	60.1	66.8
Ours		**88.8**	**68.4**	**46.1**	**56.2**	**70.6**	**87.9**	**62.4**	**88.1**	**62.2**	**67.0**
Avg.											
Continual FT	25.5	81.5	59.1	53.2	64.7	51.8	63.2	64.3	69.7	31.8	49.7
LwF [17]	36.3	86.9	72.0	59.0	73.7	60.0	73.6	74.8	80.0	37.3	58.1
iCaRL [27]	35.5	89.2	72.2	60.6	68.8	70.0	78.2	62.3	81.8	41.2	62.5
LwF-VR [4]	29.6	87.7	74.4	59.5	72.4	63.6	77.0	66.7	81.2	43.7	60.7
WiSE-FT [34]	26.7	86.5	64.3	57.1	65.7	58.7	71.1	70.5	75.8	36.9	54.6
ZSCL [46]	45.1	92.0	80.1	64.3	79.5	81.6	89.6	75.2	88.9	64.7	68.0
Ours	**46.6**	**93.4**	**82.0**	**67.4**	**82.6**	**83.7**	**90.0**	**75.7**	**89.9**	**66.8**	**68.4**
Last											
Continual FT	31.0	89.3	65.8	67.3	88.9	71.1	85.6	**99.6**	92.9	77.3	81.1
LwF [17]	26.3	87.5	71.9	66.6	79.9	66.9	83.8	**99.6**	92.1	66.1	80.4
iCaRL [27]	35.8	**93.0**	77.0	70.2	83.3	88.5	90.4	86.7	93.2	81.2	81.9
LwF-VR [4]	20.5	89.8	72.3	67.6	85.5	73.8	85.7	**99.6**	93.1	73.3	80.9
WiSE-FT [34]	27.2	90.8	68.0	68.9	86.9	74.0	87.6	**99.6**	92.6	77.8	81.3
ZSCL [46]	40.6	92.2	81.3	70.5	94.8	90.5	91.9	98.7	**93.9**	85.3	80.2
Ours	**42.4**	92.7	**83.2**	**73.2**	**97.0**	**91.8**	**92.2**	99.1	**93.9**	**87.4**	**82.6**

Transfer performance. This result shows the superiority of our method in preserving the zero-shot performance of the CLIP model while achieving preferable fine-tuning performance during continual learning. The results obtained under task order settings II demonstrate a similar trend.

Table 2 provides more details about the performance of each method on each task under task order setting I. It presents performance of final trained models on each task in the MTIL benchmark. In terms of both the Transfer and the Avg. performance, the proposed method achieved consistent performance improvement over existing baselines on all tasks. Even under the Last metric, our method also achieves competitive performance, with only a marginal gap on two datasets. More results on the MTIL benchmark can be found in the supplementary material.

Table 3: Impact of the 1st-order and the 2nd-order proximity.

		Order I			Order II		
1st-order	2nd-order	Transfer	Avg.	Last	Transfer	Avg.	Last
✗	✗	44.6±0.3	55.9±0.2	77.3±0.4	46.6±0.3	56.2±0.3	67.4±0.6
✓	✗	68.9±0.2	76.2±0.3	84.5±0.4	63.8±0.2	75.1±0.2	84.2±0.3
✗	✓	67.6±0.3	73.9±0.1	83.4±0.5	62.9±0.3	73.0±0.4	81.6±0.5
✓	✓	**69.8**±0.1	**76.9**±0.1	**85.1**±0.2	**69.8**±0.1	**76.9**±0.2	**85.1**±0.2

Table 4: Comparison of different weighting schemes.

	Order I			Order II		
Scheme	Transfer	Avg.	Last	Transfer	Avg.	Last
Only C_0	69.3±0.2	76.4±0.3	84.0±0.4	65.2±0.2	75.6±0.3	84.7±0.5
Only C_{i-1}	69.2±0.1	76.7±0.1	84.8±0.5	65.1±0.2	75.8±0.2	85.2±0.5
Average weighting	69.1±0.1	76.6±0.2	84.6±0.5	65.0±0.1	75.5±0.3	84.8±0.4
Adaptive weighting	**69.8**±0.1	**76.9**±0.1	**85.1**±0.2	**65.4**±0.1	**75.9**±0.2	**85.4**±0.2

5.3 Ablation study

Impact of proposed proximities. We conduct experiments on the MTIL benchmark to study the impact of these the 1st-order and the 2nd-order proximities on continental learning performance. We perform 5 experiments and report the mean and standard deviation of the results in Table 3. The results indicate that the model exhibits the poorest performance across all three metrics when neither proximity is preserved. Introducing either the 1st-order or the 2nd-order proximity independently enhances the model's performance. Furthermore, the best result is achieved when both the two proximities are adopted, with the maximum performance improvements being 25.2% for Transfer, 21.0% for Avg., and 7.8% for Last metrics.

Adaptive weighting. To assess the effectiveness of the proposed sample reweighting mechanism, we conduct experiments to compare it with several other weighting schemes on the MTIL benchmark, which are 1-0 weighting (Only C_0), 0-1 weighting (Only C_{i-1}), and average weighting. Table 4 illustrates the results. With the proposed method using C_{i-1} as the teacher only incurs a slight decline in the Transfer performance, but achieving remarkable performance improvement under the other two metrics, compared to using C_0 as the teacher. This demonstrates the effectiveness of the proposed method in preserving the original structure of out-domain feature space. Combining both teacher C_0 and C_{i-1} by average weighting further leads to improved downstream performance, while it also suffers from lower zero-shot performance. In contrast, adopting the proposed adaptive weighting scheme achieves competitive Avg. and Last performance and significantly improves the Transfer performance. The experiment results demonstrate that our adaptive weighting scheme effectively balances the stability and plasticity of the CLIP model during continual learning.

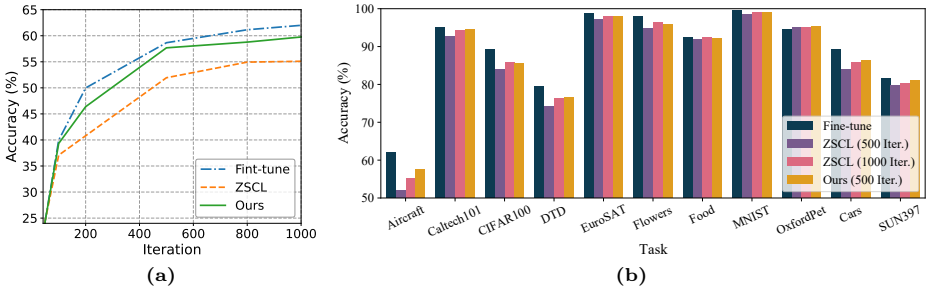

Fig. 2: Convergence performance. (a) Convergence performance of ZSCL and our proposed method on Aircraft. Result achieved by Fine-tune performs as the upper bound. (b) Immediate accuracy measured after training on each task during continual learning.

Table 5: CIL results on CIFAR100 dataset.

	10 steps		20 steps		50 steps	
Method	**Avg.**	**Last**	**Avg.**	**Last**	**Avg.**	**Last**
UCIR [11]	58.66	43.39	58.17	40.63	56.86	37.09
DER [36]	74.64	64.35	73.98	62.55	72.05	59.76
DyTox+ [6]	74.10	62.34	71.62	57.43	68.90	51.09
Zero-shot	74.57	65.98	75.34	65.98	75.88	65.98
Continual FT	65.46	53.23	59.69	43.13	39.23	18.89
LwF [17]	65.86	48.04	60.64	40.56	47.69	32.90
iCaRL [27]	79.35	70.97	73.32	64.55	71.28	59.07
LwF-VR [4]	78.81	70.75	74.54	63.54	71.02	59.45
Cont.-CLIP [32]	75.17	66.72	75.95	66.72	76.49	66.72
ZSCL [25]	82.15	73.65	80.39	69.58	79.92	67.36
Ours	**86.15**	**79.75**	**85.38**	**77.68**	**81.91**	**70.17**

5.4 Convergence performance

We evaluate the accuracy of the model at different training iterations on Aircraft. As illustrated in Figure 2a, our method exhibits significantly faster convergence compared to ZSCL. Particularly, the accuracy of our proposed model after training for 500 iterations is higher than that achieved by ZSCL after 1000 iterations.

We also present the immediate accuracy on each task during continual learning in Figure 2b, where the accuracy on each task is measured after completing the training on this task. Similarly, our proposed method outperforms ZSCL across all tasks with only a half iterations. It is noteworthy that on the Oxford-Pet, our method even outperforms direct fine-tuning of the original CLIP model, which can be considerated as upper bound. This phenomenon can be attributed to capability of our method to retain knowledge from previous tasks and apply it to the learning of current tasks. The better convergence performance of the proposed method shows that our trained model is adapted to new tasks easier.

Table 6: CIL results on TinyImageNet dataset.

	5 steps		10 steps		20 steps	
Method	**Avg.**	**Last**	**Avg.**	**Last**	**Avg.**	**Last**
EWC [13]	19.01	6.00	15.82	3.79	12.35	4.73
UCIR [11]	50.30	39.42	48.58	37.29	42.84	30.85
DyTox [6]	55.58	47.23	52.26	42.79	46.18	36.21
Zero-shot	69.68	65.47	69.62	65.47	69.56	65.47
Continual FT	61.54	46.66	57.05	41.54	54.62	44.55
LwF [17]	60.97	48.77	57.60	44.00	54.79	42.26
iCaRL [27]	77.02	70.39	73.48	65.97	69.65	64.68
LwF-VR [4]	77.56	70.89	74.12	67.05	69.94	63.89
Cont.-CLIP [32]	70.49	66.43	70.55	66.43	70.51	66.43
ZSCL	80.27	73.57	78.61	71.62	77.18	68.30
Ours	**82.43**	**77.53**	**81.74**	**76.51**	**80.68**	**75.48**

5.5 Class Incremental Learning

To evaluate the proposed methods under the CIL setting, we conduct experiments on both CIFAR100 and TinyImageNet datasets. Results achieved on CIFAR100 and TinyImageNet are presented in Table 5 and Table 6, respectively. On both datasets, our method outperforms existing approaches in terms of both Avg. and Last performance. The learning task becomes more challenge with the value CIL steps increases, but the propose method still surpass the others by a significant margin. Specifically, when training on CIFAR100 with 50 steps, our method achieves 81.91% Avg. accuracy and 70.17% Last accuracy, which are 1.98% and 2.81% higher than that achieved by the second best baseline, respectively. When training on TinyImageNet with 20 steps, the gaps even increase to 3.50% and 7.18%, respectively. The results in CIL benchmark further demonstrate our method is capable of not only preserving zero-shot transfer but also retaining knowledge acquired from previous tasks, which contributes to mitigating catastrophic forgetting in continual learning.

6 Conclusion

In this paper, we present a novel strategy for improving continual learning in vision-language models while preserving zero-shot capabilities. Specifically, Our approach focuses on balancing the model's zero-shot transfer capabilities with its adaptability to new data distributions. By constructing a cross-modal graph, we explore both inter- and intra-modal proximity. A distillation paradigm is introduced to preserve these proximity from two teachers, incorporated with a strategic sample re-weighting mechanism. This method effectively mitigates the potential conflict between model stability and plasticity, allowing both of these abilities to improve simultaneously. The results of the experiment, conducted on several benchmarks, demonstrate the effectiveness of the proposed approach.

Acknowledgements

This work was supported in part by the Australian Research Council under Projects DP240101848 and FT230100549.

References

1. Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M., Tuytelaars, T.: Memory aware synapses: Learning what (not) to forget. In: Proceedings of the European conference on computer vision (ECCV). pp. 139–154 (2018) 2
2. Bai, Z., Liu, X., Hu, H., Guo, T., Zhang, Q., Wang, Y.: Data-free distillation of language model by text-to-text transfer. arXiv preprint arXiv:2311.01689 (2023) 4
3. Chen, H., Guo, T., Xu, C., Li, W., Xu, C., Xu, C., Wang, Y.: Learning student networks in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6428–6437 (2021) 4, 8
4. Ding, Y., Liu, L., Tian, C., Yang, J., Ding, H.: Don't stop learning: Towards continual learning for the clip model. arXiv preprint arXiv:2207.09248 (2022) 2, 3, 4, 10, 11, 13, 14
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 3
6. Douillard, A., Ramé, A., Couairon, G., Cord, M.: Dytox: Transformers for continual learning with dynamic token expansion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9285–9295 (2022) 9, 13, 14
7. Fernando, C., Banarse, D., Blundell, C., Zwols, Y., Ha, D., Rusu, A.A., Pritzel, A., Wierstra, D.: Pathnet: Evolution channels gradient descent in super neural networks. arXiv preprint arXiv:1701.08734 (2017) 3
8. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. International Journal of Computer Vision pp. 1–15 (2023) 3
9. Guo, T., Xu, C., He, S., Shi, B., Xu, C., Tao, D.: Robust student network learning. IEEE transactions on neural networks and learning systems **31**(7), 2455–2468 (2019) 3
10. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015) 4
11. Hou, S., Pan, X., Loy, C.C., Wang, Z., Lin, D.: Learning a unified classifier incrementally via rebalancing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 831–839 (2019) 13, 14
12. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021) 3
13. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences **114**(13), 3521–3526 (2017) 3, 14

14. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009) 9
15. Le, Y., Yang, X.: Tiny imagenet visual recognition challenge. CS 231N **7**(7), 3 (2015) 9
16. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10965–10975 (2022) 1
17. Li, Z., Hoiem, D.: Learning without forgetting. IEEE transactions on pattern analysis and machine intelligence **40**(12), 2935–2947 (2017) 3, 10, 11, 13, 14
18. Liu, Y., Cao, J., Li, B., Yuan, C., Hu, W., Li, Y., Duan, Y.: Knowledge distillation via instance relationship graph. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7096–7104 (2019) 4
19. Lopes, R.G., Fenu, S., Starner, T.: Data-free knowledge distillation for deep neural networks. arXiv preprint arXiv:1710.07535 (2017) 4
20. Lopez-Paz, D., Ranzato, M.: Gradient episodic memory for continual learning. Advances in neural information processing systems **30** (2017) 2
21. Mallya, A., Lazebnik, S.: Packnet: Adding multiple tasks to a single network by iterative pruning. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 7765–7773 (2018) 3
22. Masana, M., Liu, X., Twardowski, B., Menta, M., Bagdanov, A.D., Van De Weijer, J.: Class-incremental learning: survey and performance evaluation on image classification. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(5), 5513–5533 (2022) 9
23. Nie, Y., He, W., Han, K., Tang, Y., Guo, T., Du, F., Wang, Y.: Lightclip: Learning multi-level interaction for lightweight vision-language models. arXiv preprint arXiv:2312.00674 (2023) 1
24. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3967–3976 (2019) 4
25. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 1, 2, 3, 4, 10, 13
26. Rannen, A., Aljundi, R., Blaschko, M.B., Tuytelaars, T.: Encoder based lifelong learning. In: Proceedings of the IEEE international conference on computer vision. pp. 1320–1328 (2017) 2
27. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: Incremental classifier and representation learning. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 2001–2010 (2017) 2, 10, 11, 13, 14
28. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550 (2014) 4
29. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022) 1
30. Shmelkov, K., Schmid, C., Alahari, K.: Incremental learning of object detectors without catastrophic forgetting. In: Proceedings of the IEEE international conference on computer vision. pp. 3400–3409 (2017) 2

31. Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., Mei, Q.: Line: Large-scale information network embedding. In: Proceedings of the 24th international conference on world wide web. pp. 1067–1077 (2015) 6
32. Thengane, V., Khan, S., Hayat, M., Khan, F.: Clip model is an efficient continual learner. arXiv:2210.03114 (2022) 3, 10, 13, 14
33. Wang, R., Duan, X., Kang, G., Liu, J., Lin, S., Xu, S., Lü, J., Zhang, B.: Attriclip: A non-incremental learner for incremental knowledge learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3654–3663 (2023) 3
34. Wortsman, M., Ilharco, G., Kim, J.W., Li, M., Kornblith, S., Roelofs, R., Lopes, R.G., Hajishirzi, H., Farhadi, A., Namkoong, H., et al.: Robust fine-tuning of zero-shot models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7959–7971 (2022) 10, 11
35. Xing, Y., Wu, Q., Cheng, D., Zhang, S., Liang, G., Zhang, Y.: Class-aware visual prompt tuning for vision-language pre-trained model. arXiv preprint arXiv:2208.08340 (2022) 1
36. Yan, S., Xie, J., He, X.: Der: Dynamically expandable representation for class incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3014–3023 (2021) 13
37. Yao, L., Han, J., Wen, Y., Liang, X., Xu, D., Zhang, W., Li, Z., Xu, C., Xu, H.: Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. Advances in Neural Information Processing Systems **35**, 9125–9138 (2022) 1
38. Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., Xu, C.: Filip: Fine-grained interactive language-image pre-training. arXiv preprint arXiv:2111.07783 (2021) 1
39. Yim, J., Joo, D., Bae, J., Kim, J.: A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4133–4141 (2017) 4
40. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917 (2022) 1
41. Yu, Y.C., Huang, C.P., Chen, J.J., Chang, K.P., Lai, Y.H., Yang, F.E., Wang, Y.C.F.: Select and distill: Selective dual-teacher knowledge transfer for continual learning on vision-language models. arXiv preprint arXiv:2403.09296 (2024) 3
42. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928 (2016) 4
43. Zang, Y., Li, W., Zhou, K., Huang, C., Loy, C.C.: Unified vision and language prompt learning. arXiv preprint arXiv:2210.07225 (2022) 1
44. Zhang, J., Huang, J., Jin, S., Lu, S.: Vision-language models for vision tasks: A survey. arXiv preprint arXiv:2304.00685 (2023) 3
45. Zheng, Z., Ma, M., Wang, K., Qin, Z., Yue, X., You, Y.: Preventing zero-shot transfer degradation in continual learning of vision-language models. arXiv preprint arXiv:2303.06628 (2023) 2, 3, 5, 8
46. Zheng, Z., Ma, M., Wang, K., Qin, Z., Yue, X., You, Y.: Preventing zero-shot transfer degradation in continual learning of vision-language models. arXiv preprint arXiv:2303.06628 (2023) 9, 10, 11
47. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022) 3