

# CTP: Towards Vision-Language Continual Pretraining via Compatible Momentum Contrast and Topology Preservation

Hongguang Zhu<sup>1,2\*</sup> Yunchao Wei<sup>1,2,3</sup> Xiaodan Liang<sup>3,4,5</sup> Chunjie Zhang<sup>1,2</sup> Yao Zhao<sup>1,2,3†</sup>

<sup>1</sup>Institute of Information Science, Beijing Jiaotong University <sup>2</sup>Beijing Key Laboratory of Advanced Information Science and Network

<sup>3</sup>Peng Cheng Laboratory <sup>4</sup>Sun Yat-sen University <sup>5</sup>MBZUAI

## Abstract

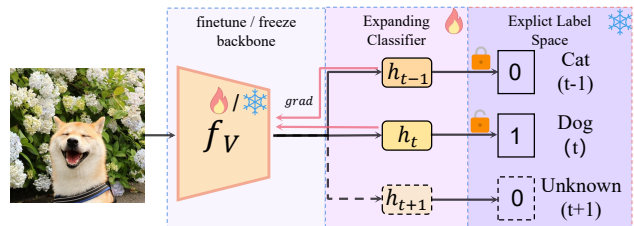
Vision-Language Pretraining (VLP) has shown impressive results on diverse downstream tasks by offline training on large-scale datasets. Regarding the growing nature of real-world data, such an offline training paradigm on ever-expanding data is unsustainable, because models lack the continual learning ability to accumulate knowledge constantly. However, most continual learning studies are limited to uni-modal classification and existing multi-modal datasets cannot simulate continual non-stationary data stream scenarios. To support the study of Vision-Language Continual Pretraining (VLCP), we first contribute a comprehensive and unified benchmark dataset P9D which contains over one million product image-text pairs from 9 industries. The data from each industry as an independent task supports continual learning and conforms to the real-world long-tail nature to simulate pretraining on web data. We comprehensively study the characteristics and challenges of VLCP, and propose a new algorithm: Compatible momentum contrast with Topology Preservation, dubbed CTP. The compatible momentum model absorbs the knowledge of the current and previous-task models to flexibly update the modal feature. Moreover, Topology Preservation transfers the knowledge of embedding across tasks while preserving the flexibility of feature adjustment. The experimental results demonstrate our method not only achieves superior performance compared with other baselines but also does not bring an expensive training burden. Dataset and codes are available at <https://github.com/KevinLight831/CTP>.

## 1. Introduction

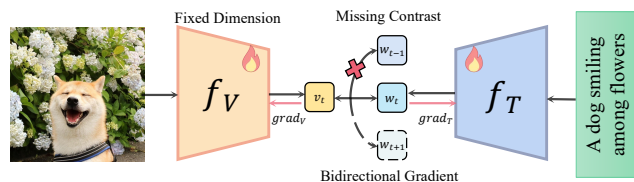
Benefiting from the remarkable generalization ability derived from large-scale pretraining, Vision-Language Pretraining (VLP) [33, 23] has emerged as the prevalent approach for addressing downstream vision-language tasks. The recent advancements in artificial intelligence such as CLIP [33] and ChatGPT [1] have further fueled this trend

\* This work was done when Hongguang Zhu worked as a research intern in Peng Cheng Laboratory. Email: kevinlight831@gmail.com

† Corresponding author: yzhao@bjtu.edu.cn



(a) Class-Incremental Learning (CIL).



(b) Vision-Language Continual Pretraining (VLCP).

Figure 1: The traditional Class-Incremental Learning (CIL) is inflexible in the continual learning of visual concepts, which needs ever-expanding classifier parameters and endless human annotation. Moreover, it is difficult for single-class labeling to cover all visual concepts in an image. *e.g.*, CIL only focuses on the foreground class dog and ignores the background class flower, while Vision-Language Continual Pretraining (VLCP) can flexibly represent the image content by text. Compared with CIL, which fixes the label space and only updates the image encoder, VLCP updates the image and text encoders simultaneously in the fixed dimension. Meanwhile, previous-task data also cannot be used as contrast samples in the continual pretraining.

of using larger models and more data. In the long term, this computational headlong rush does not seem reasonable to move toward sustainable solutions and actually also excludes academic laboratories with limited resources. Current VLP paradigms all train on prepared data in advance. Nevertheless, the world is ever-changing. Offline-trained models can not evolve in a dynamic environment to continually acquire, integrate and accumulate new knowledge. Moreover, repeated offline pretraining on the ever-expanding dataset will impose growing and endless training costs. Only finetuning on new data will also suffer severe

degradation due to *catastrophic forgetting* [32]. Hence, in practical application scenarios, it is significant for VLP to continually integrate knowledge from the incoming data.

Prior studies on continual learning [20, 55, 42, 53, 51] focused on supervised class-incremental learning (CIL), as Figure 1(a), which aims to maintain discriminative features for known classes and expand new classifiers to learn new classes. However, this paradigm is inflexible due to the constant demand for laborious annotation and increasing classifier parameters. In contrast, VLP allows for learning open-world visual patterns without explicit “class” concepts, which can capture more comprehensive visual concepts rather than just category-based features. Moreover, massive web weak-aligned image-text pairs can be used as training data without extra human annotation, and no extra parameters are needed for category expansion as the output dimension is fixed for image and text encoders.

Nevertheless, Vision-Language Continual Pretraining (VLCP) remains understudied due to the lack of datasets that satisfy both massive image-text pairs and continual tasks with discrepant knowledge. Therefore, we contribute the first VLCP dataset **P9D**, which contains more than 1 million product image-text pairs and over 3800 categories from 9 industries. Different task data are split according to industry (e.g. food and clothing) to support the continual pretraining. P9D not only is larger than previous CIL datasets both in terms of both class number and data size, but also conforms to real-world long-tailed distributions. As shown in the Figure 1(b), VLCP, as a new paradigm, also suffers new challenges compared with traditional CIL.

1) **Fixed-dimensional embedding:** CIL methods typically address the stability-plasticity dilemma by preserving old logits [27] or freezing old backbone [31] and finetuning new classifiers. Without explicit class supervision and increasing embedding dimension, the VLCP can only adjust the fixed-dimensional shared embedding to incorporate both old and new knowledge. 2) **Missing contrast samples:** CIL still can use the gradient from negative logits of old classes [27, 34, 17] even if the old data is unseen. But the lack of contrast samples from old tasks leads to sub-optimal shared embedding in VLCP. 3) **Multi-modal/task optimization:** Unlike CIL has fixed label space to optimize image encoder, VLCP involves the complicated joint optimization of image, text, and multi-modal encoders.

Therefore, we propose a simple yet effective method, Compatible momentum contrast with Topology Preservation (CTP), which maintains a compatible momentum model that absorbs both new and old knowledge to separately adjust uni-modal and multi-modal encoders. Moreover, different from CIL methods that distill visual features across tasks, topology preservation keeps consistent sample relationship across tasks. It not only transfers the topology knowledge of the old embedding while preserving the

flexibility of feature adjustment. Meanwhile, to systematically investigate the vision-language continual pretraining, we extend a series of traditional CIL methods to VLCP and evaluate them in a unified setting. Interestingly, we find that the multi-modal fusion feature by masked modeling pretraining has a strong anti-forgetting ability, and the performance of continual finetuning approximates that of joint training in multi-modal retrieval. Oppositely, due to the lack of contrastive samples from different tasks, the cross-modal alignment ability suffers serious forgetting in continual pretraining and has a big gap with joint training. Meanwhile, The experimental results show our method is not only compatible with both memory-buffer and memory-free situations, but also achieve leading performance without incurring expensive training time costs.

**Our contributions** are as follows:

- We build the first Vision-Language Continual Pretraining (VLCP) benchmark dataset P9D to support the study of VLCP, which contains massive image-text pairs and the continual non-stationary task data.
- We systematically study the characteristics and challenges of VLCP, and establish baseline library for VLCP by extending popular continual learning methods and evaluating them in a unified setting.
- We propose a simple yet effective method CTP for VLCP, which achieve both superior performance and efficient training.

## 2. Related Work

### 2.1. Vision-language pretraining

Vision-Language Pretraining (VLP) [15] leverages large-scale web image-text pairs as pretraining data and adopts self-supervised learning (contrastive learning [33] or masked modeling [4]) to train the transferable image-text embeddings. The VLP models can coarsely be divided into two paradigms according to architectures: 1) Dual-Encoder and 2) Fusion-Encoder. **Dual-Encoder** models encode images and texts respectively by separate encoders and employ cosine similarity to build the image-text alignment. The Dual-Encoder models [29, 33, 19] achieve promising results on image-text retrieval with linear time complexity. However, the loose modal interaction by cosine similarity also limits the multi-modal fusion ability [3, 9, 46]. Thus, the other paradigm **Fusion-Encoder** employs cross-modal attention to jointly encode images and text. The prior works [8, 25, 52] use pretrained detectors to extract regional features and Transformers [40] for multi-modal fusion. However, extracting region features is computationally expensive and the joint transformer requires quadratic time complexity for retrieval tasks. Thus, align before fuse (ALBEF) architecture [24, 23, 48] incorporates the image-text contrastive loss before multi-modal fusion and replaces

the detector with ViT [12] for the end-to-end pretraining. It not only eliminates the burden of object detection preprocessing but also keeps multi-modal interaction by the top fusion layers and linear retrieval time complexity by the bottom dual encoders.

Nevertheless, current VLP models are all trained in a joint manner using prepared data and keep fixed pretrained parameters. In the long term, they cannot constantly accumulate knowledge and evolve themselves to accommodate the dynamic world. Thus, we concentrate on the vision-language continual pretraining based on the ALBEF architecture and evaluate the cross-modal alignment and multi-modal fusion capabilities in the continual environment.

## 2.2. Continual Learning

Continual learning aims to overcome catastrophic forgetting and integrate novel knowledge in a sequential fashion where old data are unavailable. Conventional continual learning methods mainly focus on image classification tasks. They can be roughly categorized into three groups: 1) **Regularization-based methods** [20, 2, 49, 41, 6, 27] limit the plasticity of the model to address catastrophic forgetting by regularizing important parameters or knowledge distillation. Although these methods alleviate forgetting to some extent without storing old samples, they cannot get satisfactory performance in some complex datasets [45] and challenging settings [30]. 2) **Architecture-based methods** [35, 36, 31, 26] keep old parameters fixed while growing and allocating weights for learning new data. These methods can expand sub-networks or focus more on the specific part of network modules. However, these models require task identity to condition the network at test time, which is impractical for more realistic and task-agnostic settings *e.g.* retrieval tasks. Additionally, as the number of tasks increases, the parameters of the added sub-networks become very huge, which is also not suitable for application deployment. 3) **Replay-based methods** [7, 34, 17] apply extra memory to store a few samples from previous tasks or learn to generate pseudo data and train with the current data together. Based on this simple yet effective idea, recent methods further improve and achieve state-of-the-art performance by involving different sampling strategies. However, the memory size and the training complexity will be enlarged significantly and unaffordable as the growth of tasks, especially for costly large-scale pretraining.

However, It is an inflexible way to get the foundation model with continual learning ability through image classification. Firstly, most real-world data can not be accurately represented semantics by simple category labeling, and class annotation is labor-intensive. Secondly, the ever-expanding classifier will also bring endless growth of parameters. In contrast, VLP does not require explicit "class" labeling and can cover wider visual concepts by text, while

its fixed output dimensions can continually support downstream tasks without increasing parameters.

## 3. Dataset

Traditional CIL datasets [22, 21] usually use simple images with single-semantic and have the same sample number for different tasks. This ideal setting is difficult to simulate the real-world data which is noisy and long-tail distributed. Meanwhile, despite massive web data collected by some VLP datasets [37, 5], The simple random partitioning will make each chunk still conform to the overall distribution and unable to simulate the continual environment. Considering the rich product samples in e-commerce websites, which not only conform to the real web-data nature but also have category, industry, and title information as weakly semantic correspondence, we use e-commerce data to construct the first vision-language continual pretraining dataset P9D and establish the unified evaluation benchmark.

### 3.1. Dataset Split

As the Figure 2(c), P9D includes more than 1 million image-text pairs of real products. According to the industry name of products, P9D is divided into 9 tasks to sequential training (default order) which are Household, Furnishings, Food, Beauty, Clothing, Auto, Parenting, Outdoor, and Electronics. We select 1,014,599 image-text pairs for training, and 2,846 pairs as the test set of cross-modal retrieval, 4,615 and 46,855 pairs as the query set and gallery set of multi-modal retrieval. The quantity distributions of the four sets are consistent and more details can refer to the appendix. Considering the descriptions of similar products may be very similar, to avoid the situation that one image corresponds to multiple captions affecting the overall evaluation of cross-modal retrieval, we save one sample for the categories owning more than 100 samples in the training set. Meanwhile, For the query set, the number of samples per class is about 0.5% of that in the training set.

### 3.2. Data Analysis.

**Real-world Web Data:** We collect massive commodity data from e-commerce websites and split task identity according to industry to simulate real-world rich-concept but non-stationary data streams. Different from the existing CIL datasets with clear images and labels, our images present the characteristics of multi-domain mixing (as Figure 2(b)). *e.g.*, multiple backgrounds, amorphous watermarks, occlusion, and multi-view. Meanwhile, although only text can be used in training, P9D also includes product class labels to evaluate the fusion feature clustering by multi-modal retrieval.

**Rich Categories:** Some recent papers can obtain preferable classification results on traditional CIL datasets by finetun-

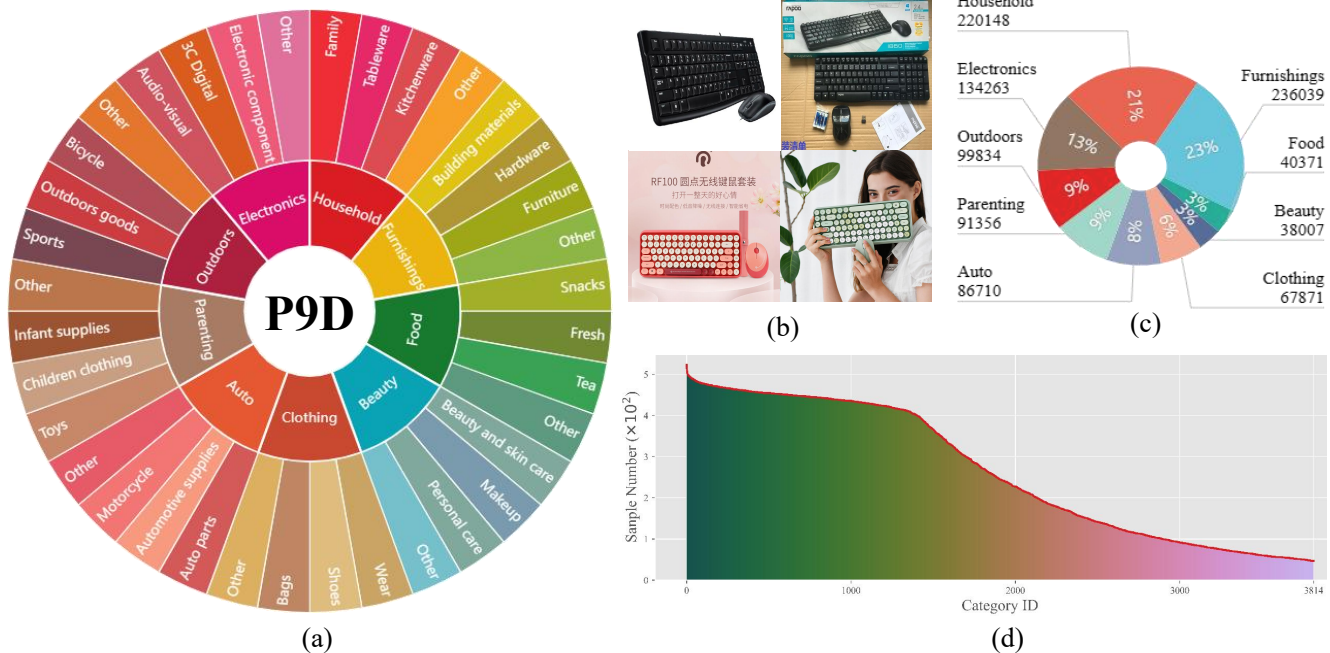


Figure 2: The dataset statistic of P9D. (a) The abbreviated hierarchical categories structure of P9D. (b) Some multi-domain mixing examples about the product ‘Keyboard and mouse set’. (c) The quantity distribution of P9D train set. (d) The sample number distribution of category in P9D. The red line represents the sample number of each category in decreasing order.

ing [43] and even freezing [38] the weight of pretrained models. We suppose that the traditional benchmark datasets contain limited classes (e.g. MNIST [22] and CIFAR-100 [21]) and thus cannot adequately evaluate the continual learning methods in the face of powerful generalized pretrained models. Unlike these datasets with the narrow space of category labels, our dataset contains over 3800 categories and divides each task in a way that matches the real-world industry domains. Therefore, our setting is more challenging and realistic. The abbreviated hierarchical categories structure of P9D is shown in the Figure 2(a), and more comparisons with other datasets can refer to the appendix.

**Real-World Distribution:** The conventional CIL datasets consider a balanced distribution for each task but ignore the nature of long-tailed distributions in the real world. In contrast, our dataset contains a different number of categories for each task and the long-tailed distribution aligns well with real-world scenarios. The Figure 2(d) shows the sample number distribution of the categories.

## 4. Methodology

### 4.1. Preliminary

**Problem Setting:** We propose a vision-language continual pretraining (VLCP) setting, where models are supposed to be sequentially trained on  $T$  tasks data  $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_T\}$ . In the  $t$ -th task, the whole sub-dataset

$\mathcal{D}_t = \{(I_i^t, T_i^t)\}_{i=1}^{N_t}$  contains the  $N_t$  image-text pairs where  $I_i^t$  and  $T_i^t$  respectively denote the  $i$ -th image and the corresponding text description of the  $t$ -th task. Because the old task data is unseen in the continual setting, the VLP model trained on the current dataset  $\mathcal{D}_t$  needs to resist forgetting and performs well for all learned datasets  $\{\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_t\}$ .

**Model Architecture:** As shown in Figure 3, We use a 12-layer ViT-B/16 [12] as the image encoder  $f_v$ , and initialize it with weights pretrained on ImageNet-1k from [39]. The image  $I$  is encoded into patch feature sequence  $f_v(I) = \{v^{\text{cls}}, v^1, \dots, v^N\}$ . Meanwhile, We use the first and the last 6 layers of BERT<sub>base</sub> [10] model to initialize the text encoder  $f_t$  and multi-modal encoder  $f_m$ . The text  $T$  is encoded into word feature sequence  $f_t(T) = \{w^{\text{cls}}, w^1, \dots, w^N\}$ . Then, the text features  $f_t(T)$  will fuse with the image features  $f_v(I)$  through cross attention at each layer of multi-modal encoder  $f_m$ . Two linear transformations  $g_v$  and  $g_t$  are map the  $v^{\text{cls}}$  and  $w^{\text{cls}}$  to the low-dimensional (256-d) representations for Image-Text Alignment (ITA). It can be denoted  $v = g_v(v^{\text{cls}})$  and  $w = g_w(w^{\text{cls}})$ .

Given a pair of image-text data  $(I_i, T_j)$  in the current  $\mathcal{D}_t$ , the image-text similarity function  $s(I_i, T_j)$  is defined as the cosine similarity  $s(I_i, T_j) = \frac{v_i^T w_j}{\|v_i\| \|w_j\|}$ . Given a batch of  $B$  pairs, the model uses the symmetric cross-entropy loss over the  $B \times B$  similarity matrix to optimize the parameters. The image-to-text loss  $\mathcal{L}_{i2t}$  and the text-to-image loss  $\mathcal{L}_{t2i}$  are

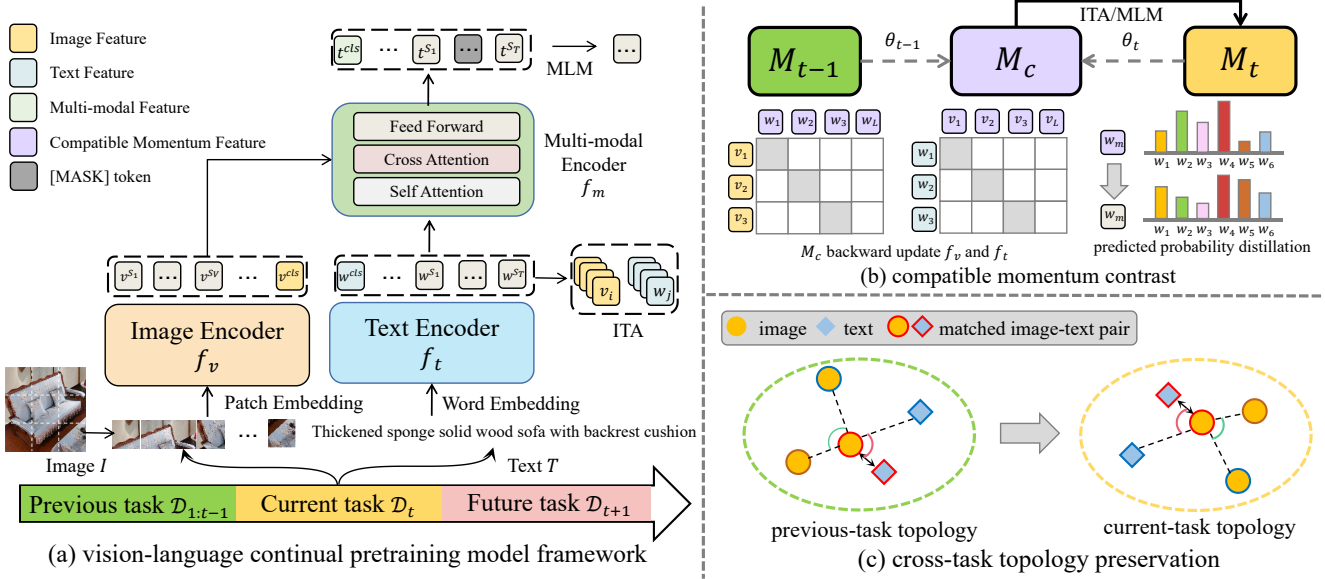


Figure 3: Illustration of the proposed vision-language continual pretraining method CTP. (a) shows the overall continual pretraining model architecture, and (b) shows the interactive adjustment of the compatible momentum model  $M_c$  and current training model  $M_t$ . The  $M_c$  absorb the parameter of current model  $M_t$  and previous-task model  $M_{t-1}$ , and in turn constrains the update of  $M_t$ . (c) shows that the topology relationship of previous task is maintained as much as possible while allowing the overall embedding to be updated.

formulated as:

$$\begin{aligned} \mathcal{L}_{i2t} &= -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(s(I_i, T_i)/\tau)}{\sum_{j=1}^B \exp(s(I_i, T_j)/\tau)}, \\ \mathcal{L}_{t2i} &= -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(s(T_i, I_i)/\tau)}{\sum_{j=1}^B \exp(s(T_i, I_j)/\tau)}, \end{aligned} \quad (1)$$

where  $\tau$  is the temperature parameter. The Image-Text Alignment (ITA) loss is defined as:  $\mathcal{L}_{ita} = \frac{\mathcal{L}_{i2t} + \mathcal{L}_{t2i}}{2}$ .

In Masked Language Modeling (MLM), give an image-text pair, we randomly mask out the words with probability of 15% [10, 24], and replace masked ones  $w_m$  with the special token [MASK]. The goal is to predict these masked tokens based on their surrounding words and the image features, by minimizing the cross-entropy loss:

$$\mathcal{L}_{mlm} = -\mathbb{E}_{(I, \hat{T}) \sim D_t} \mathbb{H}(y^m, p_\theta^m(I, \hat{T})), \quad (2)$$

where the  $y^m$  is the one-hot vocabulary distribution where the masked token has a probability of 1, and  $p_\theta^m(I, \hat{T})$  is the predicted probability of model  $\theta$  for masked token  $w_m$ . The total VLP loss is defined as  $\mathcal{L}_{VLP} = \mathcal{L}_{ita} + \mathcal{L}_{mlm}$ .

## 4.2. Compatible Momentum Contrast

Due to the fixed-dimensional embedding, VLP cannot isolate old and new knowledge like CIL by extending projection parameters. Therefore, both the vision and language embeddings need to be constantly adjusted to simultaneously accommodate the old and new image-text pairs.

In order to review old knowledge and adapt to the new task, we use the momentum model  $M_c$  initialized by the previous-task trained model  $M_{t-1}$  as additional supervision of the current training model  $M_t$ . Some single-modal [16] or vision-language [18] pretraining works also adopt momentum model as a temporal ensembling method [14] to smoothly guide the training. However, the traditional momentum model is updated by only parameters of training model. With the accumulation of training steps, it will be gradually affected by the new model and also suffers catastrophic forgetting. Moreover, recklessly maintaining old knowledge will also make the model lose the plasticity to acquire new knowledge. Therefore, we propose the compatible momentum update which synchronously absorbs the old and new knowledge:

$$\theta_c \leftarrow m \cdot \theta_c + \frac{1-m}{2} \cdot \theta_{t-1} + \frac{1-m}{2} \cdot \theta_t, \quad (3)$$

where  $m \in [0, 1)$  is the momentum coefficient and  $\theta$  is the model parameters. The adjustment of models is interactive. Compatible momentum model  $M_c$  updates parameters  $\theta_c$  through the previous-task model  $M_{t-1}$  and training model  $M_t$ . In turn, the training model  $M_t$  are optimized by the back-propagation of momentum contrast and affects the parameters  $\theta_t$  to be passed in the next step. To further steadily update uni-modal encoders  $f_v$  and  $f_t$ , we maintain two dynamic queues to preserve the  $K$  negative image/text features. The image features  $v_i^c$  and text feature  $w_i^c$  from compatible momentum encoders are constantly

pushed into visual queue  $Q^I = \{v_1^c, v_2^c, \dots, v_{N_Q}^c\}$  and text queue  $Q^T = \{w_1^c, w_2^c, \dots, w_{N_Q}^c\}$  which  $N_Q = K + B$ . The compatible momentum contrastive losses about the vision and language encoders can be formulated as follows:

$$\begin{aligned}\mathcal{L}_{i2t}^c &= -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(s(I_i, Q_i^T)/\tau)}{\sum_{j=1}^{N_Q} \exp(s(I_i, Q_j^T)/\tau)}, \\ \mathcal{L}_{t2i}^c &= -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(s(T_i, Q_i^I)/\tau)}{\sum_{j=1}^{N_Q} \exp(s(T_i, Q_j^I)/\tau)}.\end{aligned}\quad (4)$$

Similarly,  $\mathcal{L}_{ita}^c = \frac{\mathcal{L}_{i2t}^c + \mathcal{L}_{t2i}^c}{2}$ .  $\mathcal{L}_{ita}^c$  constraints that the contrastive relation of image-text pairs is still workable between the encoders of the compatible momentum model and current model. It allows slow adjustment of uni-modal encoders. Besides, The compatible momentum model also provides the soft predicted probability for masked language modeling loss.

$$\mathcal{L}_{mlm}^c = -\mathbb{E}_{(I, \hat{T}) \sim D_t} H(p_{\theta_c}^m(I, \hat{T}), p_{\theta_t}^m(I, \hat{T})), \quad (5)$$

Thus, Compatible Momentum Contrastive loss can be defined as  $\mathcal{L}_{CMC} = \mathcal{L}_{ita}^c + \mathcal{L}_{mlm}^c$ .

### 4.3. Topology Preservation

Although the compatible momentum contrast can flexibly adjust the output feature of uni-modal and fusion encoder, It does not directly transfer relationship knowledge of samples across tasks and the model may forget the overall topology of prior embedding to obtain sub-optimal performance. Unlike CIL can receive the gradient of old classes according to labels, VLCP has no gradient from old contrast samples and the image-text encoder is bidirectional synchronization adjustment according to the image-text similarity. To integrate the old and new knowledge while maintaining the topology relationship of prior tasks, we constrained the mini-batch sample relationships of the current and previous-task models to be consistent. Specifically, it is mainly divided into cross-modal topology preservation loss  $\mathcal{L}_c$  and same-modal topology preservation loss  $\mathcal{L}_s$ . To image  $I = \{I_1, I_2, \dots, I_B\}$  and text  $T = \{T_1, T_2, \dots, T_B\}$ , we define the image-to-text similarity distribution as follows:

$$\mathcal{P}^{i2t} = \frac{\exp(s(I, T_i)/\tau)}{\sum_{i=1}^B \exp(s(I, T_i)/\tau)}, \quad (6)$$

and the definition of text-to-image similarity distribution  $\mathcal{P}^{t2i}$  is similar. Thus, the  $\mathcal{L}_c$  can be formulated as the cross-entropy  $H$  between current model  $M_t$  and previous-task model  $M_{t-1}$ :

$$\mathcal{L}_c = \frac{1}{2} \mathbb{E}_{(I, T) \sim D_t} [H(\mathcal{P}_{\theta_{t-1}}^{i2t}, \mathcal{P}_{\theta_t}^{i2t}) + H(\mathcal{P}_{\theta_{t-1}}^{t2i}, \mathcal{P}_{\theta_t}^{t2i})], \quad (7)$$

In the same-modal topology preservation, the similarity of the same sample is 1 under both old and new models.

However, such a large similarity would suppress the relational distillation of other unmatched pairs [54]. Thus, we conduct the simple strategy that changes the similarity at the diagonal of  $s(I, I)$  from 1 to a minimum (*e.g.*-1000) to exclude the ‘‘apical dominance’’ of matched samples. We formulate the changed matrix as  $\hat{s}(I, I)$  and the  $\mathcal{L}_s$  can be formulated as:

$$\hat{\mathcal{P}}^{i2i} = \frac{\exp(\hat{s}(I, I_i)/\tau)}{\sum_{i=1}^B \exp(\hat{s}(I, I_i)/\tau)}, \quad (8)$$

$$\mathcal{L}_s = \frac{1}{2} \mathbb{E}_{(I, T) \sim D_t} [H(\hat{\mathcal{P}}_{\theta_{t-1}}^{i2i}, \hat{\mathcal{P}}_{\theta_t}^{i2i}) + H(\hat{\mathcal{P}}_{\theta_{t-1}}^{t2t}, \hat{\mathcal{P}}_{\theta_t}^{t2t})], \quad (9)$$

Thus, the overall continual pretraining loss is as follows:

$$\mathcal{L} = \mathcal{L}_{VLP} + \mathcal{L}_{CMC} + \mathcal{L}_c + \mathcal{L}_s, \quad (10)$$

## 5. Experiment

### 5.1. Experimental Setup

**Implementation Details** Regarding offline pretrained VLP models have established widespread generalization, direct finetuning may involve knowledge leakage and weaken the accurate evaluation of continual learning ability. Therefore, we do not load the weight of the pretrained VLP model for initialization. For each task, All models are trained for 5 epochs on 4 NVIDIA A100 GPUs with batch size 128 per GPU. We use the AdamW [28] optimizer with a weight decay of 0.05, and the learning rate is set as 1e-4 and decays to 1e-6 following a cosine schedule. We take random image crops of resolution  $224 \times 224$  during pretraining and also apply RandAugment. The maximum sequence length of tokens is limited to 30. The momentum parameter is set as 0.9, and the queue size  $K$  is set as 1024. The cross-modal alignment temperature  $\tau = 0.07$ .

**Evaluation Setting.** With the knowledge accumulation in continual learning, the evaluation galleries of cross-modal and multi-modal retrieval also need to expand by merging the data of new tasks. Similar to the standard vision-language modeling setting[33, 24, 23], For cross-modal retrieval, we measure the performance with Recall at  $K$  ( $R@K$ ,  $K=1,5,10$ ) [13], which is defined as the proportion of ground truth being retrieved at top- $K$  of the ranking list. Also, we use Rm as the overall metric, which is defined as the mean of  $R@K$  of both text and image retrieval.

The multi-modal encoder can fuse the multi-modal information and produce a comprehensive representation to predict masked words. Thus, we use the multi-modal output [CLS] token to retrieve samples of the same class to evaluate the zero-sample clustering ability of multi-modal fusion features for similar products. It will not introduce extra interference from finetuning on downstream tasks. (*e.g.*

| Methods              | Training hours | Cross-modal Retrieval |              |              |              |              |              |              | Multi-modal Retrieval |       |        |
|----------------------|----------------|-----------------------|--------------|--------------|--------------|--------------|--------------|--------------|-----------------------|-------|--------|
|                      |                | TR@1                  | TR@5         | TR@10        | IR@1         | IR@5         | IR@10        | Rm           | mAP@1                 | mAP@5 | mAP@10 |
| JointT               | –              | 61.31                 | 87.17        | 91.67        | 61.60        | 86.79        | 91.95        | 80.08        | 63.79                 | 70.10 | 67.40  |
| <b>Memory-Free</b>   |                |                       |              |              |              |              |              |              |                       |       |        |
| SeqF                 | 3.4            | 34.79                 | 62.83        | 72.03        | 35.73        | 63.63        | 72.14        | 56.86        | 62.15                 | 68.03 | 65.30  |
| SI [49]              | 4.9            | 34.96                 | 63.60        | 73.01        | 35.84        | 63.49        | 73.44        | 57.39        | 61.32                 | 67.48 | 64.78  |
| MAS [2]              | 4.1            | 36.65                 | 64.69        | 73.93        | 37.25        | 64.55        | 74.39        | 58.57        | 62.62                 | 68.63 | 65.79  |
| EWC [20]             | 4.6            | 37.28                 | 65.11        | 74.38        | 38.05        | 65.57        | 75.05        | 59.24        | 62.99                 | 68.75 | 65.62  |
| AFEC [41]            | 8.7            | 37.67                 | 66.34        | 74.67        | 38.79        | 66.48        | 75.16        | 59.85        | 62.58                 | 68.44 | 65.66  |
| LWF [27]             | 4.0            | 37.63                 | 66.55        | 75.09        | 38.26        | 66.62        | 75.26        | 59.90        | 62.34                 | 68.64 | 65.94  |
| RWalk [6]            | 6.2            | 37.77                 | 68.10        | 76.70        | 38.83        | 67.39        | 76.74        | 60.92        | 62.41                 | 68.43 | 65.59  |
| Our:CTP              | 4.0            | <b>43.43</b>          | <b>72.10</b> | <b>80.08</b> | <b>43.39</b> | <b>71.15</b> | <b>79.06</b> | <b>64.87</b> | 62.64                 | 68.21 | 65.10  |
| <b>Memory-Buffer</b> |                |                       |              |              |              |              |              |              |                       |       |        |
| MoF [34]             | 4.8            | 42.87                 | 71.93        | 80.60        | 43.71        | 72.03        | 80.67        | 65.30        | 62.84                 | 68.79 | 65.98  |
| LUCIR [17]           | 5.5            | 43.82                 | 73.68        | 82.04        | 44.38        | 73.54        | 80.99        | 66.41        | 61.06                 | 67.57 | 64.74  |
| ER [7]               | 4.4            | 45.19                 | 73.40        | 81.97        | 44.97        | 72.70        | 81.10        | 66.56        | 62.21                 | 68.39 | 65.80  |
| Kmeans [7]           | 4.7            | 46.17                 | 74.65        | 82.33        | 45.92        | 73.68        | 81.80        | 67.26        | 62.77                 | 68.82 | 66.04  |
| ICARL [34]           | 5.3            | 45.85                 | 74.63        | 82.57        | 46.24        | 73.23        | 81.83        | 67.39        | 63.51                 | 69.20 | 66.39  |
| Our:CTP+ER           | 4.7            | <b>50.53</b>          | <b>77.62</b> | <b>84.57</b> | <b>49.79</b> | <b>76.77</b> | <b>84.47</b> | <b>70.63</b> | 62.62                 | 68.68 | 65.81  |

Table 1: The final cross-modal and multi-modal retrieval performance comparison with different Memory-Free and Memory-Buffer continual learning baselines.

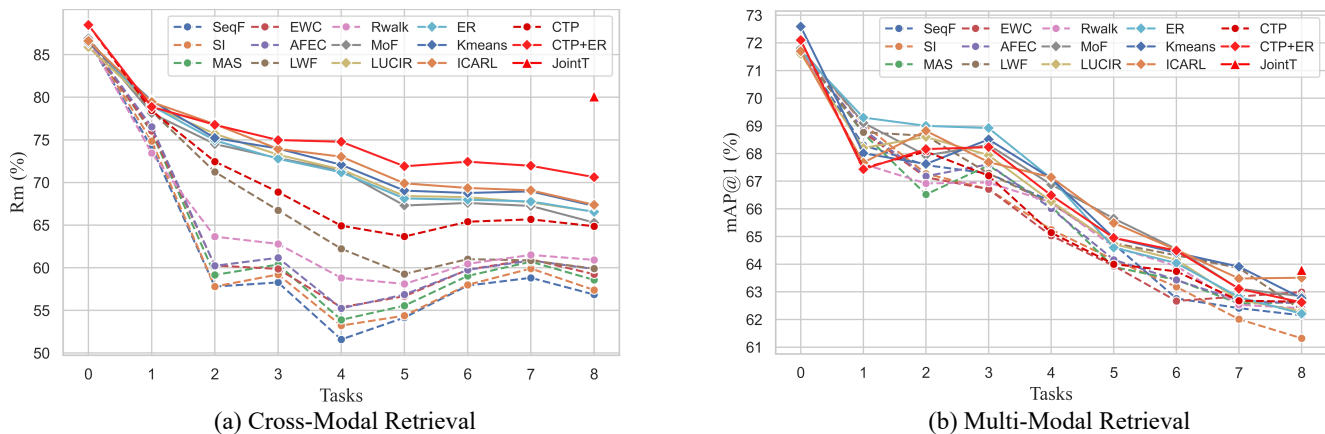


Figure 4: The performance curve of different methods on the continual learning. For Task  $i$ , the model  $\theta_i$  only test on the merged test set of learned tasks  $\{\mathcal{D}_0, \mathcal{D}_1, \dots, \mathcal{D}_i\}$ .

classification needs to train new classification head from scratch). Because each query here corresponds to multiple targets with the same class, multi-modal retrieval considers both the number and ranking of targets in the top-K retrieved candidates. For multi-modal retrieval, we adopt mean Average Precision (mAP@N) [44, 50, 11] as the evaluation metrics. mAP@N is computed as follows:

$$\text{mAP@N} = \frac{1}{Q} \sum_{i=1}^Q \frac{1}{m_q} \sum_{k=1}^N P_q(k) \delta_q(k) \quad (11)$$

$$P_q(k) = \frac{R_q(k) + 1}{k}$$

where  $Q$  is the total number of multi-modal queries and  $m_q$

is the total target number of the  $q$ -th query in the retrieved top-N relevant candidates from gallery.  $P_q(k)$  is the precision at rank  $k$  for the  $q$ -th query, and the  $\delta_q(k)$  is a binary indicator function that returns 1 when the  $k$ -th prediction is correct for the  $q$ -th query and 0 otherwise.  $R_q(k)$  is the current sum of returned correct predictions at the rank  $k$ .

**Comparison Methods.** To verify the effectiveness of our method, we compare it with several popular continual learning methods. Since all competitors are originally proposed for classification settings. We replace the instance-label cross-entropy loss with image-text contrastive loss to train the VLP model and add the Masked Language Modeling loss to train the multi-modal fusion capability as follows

[24]. Besides, We use joint training (JointT) of all seen samples as the upper-bound performance and sequential finetuning (SeqF) as the lower-bound. All baselines can be separated into Memory-Free and Memory-Buffer methods based on the availability of old samples. For the former, we evaluate representative regularization-based methods such as EWC [20], SI [49], MAS [2], AFEC [41], RWalk [6], and LwF [27]. As for the latter, we evaluate different replay sampling strategies: ER [7], Mean-of-Feature (MoF)[34], and Kmeans [7], and some replay-based methods: ICARL[34] and LUCIR [17]. We must emphasize that although Memory-Buffer methods usually have higher performance than Memory-Free methods in CIL, they bring expensive storage costs in large-scale pretraining, and performance is directly related to the replay sample size. For the comprehensive and unified study of VLCP, to all Memory-Buffer methods, we maintain a fixed-size buffer of 10000 image-text pairs (about 1% dataset) and continually update the stored samples. The detailed introduction for each method can refer to the appendix.

## 5.2. Experimental Results

**Cross-modal vs. Multi-modal.** From the Figure 4, we found an interesting phenomenon that SeqF (lower-bound) and JointT (upper bound) have a large gap (23.22% in Rm) in cross-modal retrieval, but a small gap (1.65% in mAP@1) in multi-modal retrieval. This phenomenon shows that multi-modal fusion is stronger anti-forgetting ability than cross-modal alignment in VLCP. We suppose that on the one hand, The redundancy and complementarity of multi-modal information help multi-modal fusion resist the forgetting of class attributes. On the other hand, the pre-text task MLM beforehand creates a word prediction classifier that corresponds to each word of the dictionary and keeps consistency across tasks. This predefined and well-initialized label space [47] maybe train models more stably than ever-expanding label space.

**Parameter vs. Topology Preservation.** From the comparison of Memory-Free methods in Table 1, we observe that the regularization methods [49, 2, 41, 6] represented by EWC [20] perform poorly. Probably because such methods conservatively trust the old model parameters and cannot flexibly update representation to better accommodate new knowledge. Meanwhile, They introduce the extra post-processing of calculating the fisher matrix which reduces training efficiency. In contrast, benefiting from flexible updating of compatible momentum contrast and soft constraints of topology preservation, Our CTP achieves 3.95% improvement over the most competitive method on Rm while not bring more time costs and extra memory burden.

**Memory-Free vs. Memory-Buffer.** It can be noted that the Memory-Buffer methods exhibit superior final perfor-

| $\mathcal{L}_{CMC}$ | $\mathcal{L}_c$ | $\mathcal{L}_s$ | ER | TR@1  | IR@1  | Rm    | mAP@1 |
|---------------------|-----------------|-----------------|----|-------|-------|-------|-------|
| X                   | X               | X               | X  | 34.79 | 35.73 | 56.86 | 62.15 |
| ✓                   | X               | X               | X  | 37.81 | 37.63 | 59.86 | 62.10 |
| X                   | ✓               | X               | X  | 41.45 | 40.39 | 62.50 | 61.32 |
| ✓                   | ✓               | X               | X  | 42.69 | 42.41 | 64.38 | 61.89 |
| ✓                   | X               | ✓               | X  | 41.92 | 42.02 | 63.17 | 61.11 |
| ✓                   | ✓               | ✓               | X  | 43.43 | 43.39 | 64.87 | 62.64 |
| ✓                   | ✓               | ✓               | ✓  | 50.53 | 49.79 | 70.63 | 62.62 |

Table 2: The ablation study on each component of CTP. the  $\mathcal{L}_{CMC}$  represent the compatible momentum contrast.

| Method                                | TR@1  | IR@1  | Rm    | mAP   |
|---------------------------------------|-------|-------|-------|-------|
| only $\theta^{t-1}$                   | 41.95 | 42.23 | 63.57 | 62.06 |
| only $\theta^t$                       | 40.41 | 40.34 | 62.32 | 62.32 |
| w/o $Q$                               | 42.69 | 40.83 | 63.94 | 61.63 |
| w/o $\hat{P}^{t2i}$ & $\hat{P}^{t2t}$ | 40.76 | 40.86 | 63.12 | 62.04 |
| Our:CTP                               | 43.43 | 43.39 | 64.87 | 62.64 |

Table 3: The more detailed ablation results of CTP.

mance and smaller performance fluctuations compared to the Memory-Free methods. This is attributed to the use of old data from the memory buffer as contrast samples, which plays the role of joint optimization in continual pretraining. Meanwhile, we found different replay sampling strategies have similar performance, but Kmeans[7] has a slight advantage, It may be because Kmeans can unsupervised cluster features to sample representative points of current embedding without category prior. However, the Memory-Buffer methods also bring extra storage and time cost causing by old data preservation and retraining process. The results show CTP can be further improved by 5.76% with ER sampling strategy and outperform the second-place method by 3.24% on Rm.

## 5.3. Ablations

We perform ablation of each module in the CTP method and find that each module effectively improves the cross-modal retrieval performance from the Table 2.

**CMC vs. TP.** It seems that cross-modal topology preservation plays a more direct anti-forgetting role than compatible momentum contrast in cross-modal retrieval, which improves 5.64% on Rm but also decreases 0.82% on mAP@1. When they are combined, the cross-modal and multi-modal retrieval performance were all improved by 1.88% and 0.57%. In addition, The same-modal topology preservation further improves by 0.49% and 0.75%.

**Compatible Update vs. Single-way Update.** As shown in Table 3, we compare compatible momentum update with the single-way momentum update from the current model  $\theta^t$  and the previous-step model  $\theta^{t-1}$ , and find the single-



way update can not get well results due to totally relay on the current or old model. In addition, the momentum update from  $\theta^{t-1}$  is slightly higher than that from  $\theta^t$  in Rm. This may be because the model accumulates large knowledge in continual learning, and the non-globally optimal update will lead to more forgetting in the later tasks. Therefore, it should pay more attention to the maintenance of old knowledge in the later periods of continual learning.

**Momentum Queue.:** We compare with the situation without the momentum queue ‘w/o  $Q$ ’ in Table 3 and find that the queue can slightly improve the performance. We suppose the queue stores negative samples of the current task, which plays a role of smoothing training and anti-forgetting to a certain extent.

**Suppress Same-modal Maximum Similarity.:** We find it results in a negative effect and leads to performance degradation if not suppressing same-modal maximum similarity. It indicates that the maximum similarity will obscure the sample relationship and affect the performance.

## 6. Conclusion

In this paper, we build the first Vision-Language Continual Pretraining benchmark dataset P9D which contains over 1 million image-text pairs and task data with discrepant knowledge to simulate the continual pretraining environment. Further, we comprehensively study new characteristics and challenges of VLCP, and propose a new approach CTP which combines the compatible momentum contrast and topology preservation to flexibly update model to accommodate the ever-changing embedding. It can achieve the superior performance while ensuring efficient training.

## Acknowledgement

This work was supported in part by the National Key R&D Program of China (No.2021ZD0112100), National NSF of China (No.U1936212, No.62120106009, No.62072026), Beijing Natural Science Foundation (JQ20022). We would like to acknowledge Xiao Dong and Hao Wang from Sun Yat-sen University, as well as Gengwei Zhang from University of Technology Sydney for the helpful discussions.

## References

- [1] Chatgpt. <https://openai.com/blog/chatgpt/>.
- [2] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision (ECCV)*, pages 139–154, 2018.
- [3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.
- [4] Hangbo Bao, Wenhui Wang, Li Dong, and Furu Wei. VI-beit: Generative vision-language pretraining. *arXiv preprint arXiv:2206.01127*, 2022.
- [5] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021.
- [6] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European conference on computer vision (ECCV)*, pages 532–547, 2018.
- [7] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and M Ranzato. Continual learning with tiny episodic memories. 2019.
- [8] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Proceedings of the European Conference on Computer Vision.*, 2020.
- [9] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, control and tell: A framework for generating controllable and grounded captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8307–8316, 2019.
- [10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.
- [11] Xiao Dong, Xunlin Zhan, Yangxin Wu, Yunchao Wei, Michael C Kampffmeyer, Xiaoyong Wei, Minlong Lu, Yaowei Wang, and Xiaodan Liang. M5product: Self-harmonized contrastive learning for e-commercial multi-modal pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21252–21262, 2022.
- [12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.
- [13] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *Proceedings of the British Machine Vision Conference.*, 2017.
- [14] Geoff French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. In *International Conference on Learning Representations*, 2018.
- [15] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, Jianfeng Gao, et al. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 14(3–4):163–352, 2022.

- [16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.
- [17] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 831–839, 2019.
- [18] Yuqi Huo, Manli Zhang, Guangzhen Liu, Haoyu Lu, Yizhao Gao, Guoxing Yang, Jingyuan Wen, Heng Zhang, Baogui Xu, Weihao Zheng, et al. Wenlan: Bridging vision and language by large-scale multi-modal pre-training. *arXiv preprint arXiv:2103.06561*, 2021.
- [19] 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.
- [20] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.
- [21] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- [22] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [23] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.
- [24] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. 2021.
- [25] Xiujun Li, Xi Yin, C. Li, X. Hu, Pengchuan Zhang, Lei Zhang, Longguang Wang, H. Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. *Proceedings of the European Conference on Computer Vision.*, 2020.
- [26] Xilai Li, Yingbo Zhou, Tianfu Wu, Richard Socher, and Caiming Xiong. Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting. In *International Conference on Machine Learning*, pages 3925–3934. PMLR, 2019.
- [27] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.
- [28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [29] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.
- [30] Zheda Mai, Ruiwen Li, Jihwan Jeong, David Quispe, Hyunwoo Kim, and Scott Sanner. Online continual learning in image classification: An empirical survey. *Neurocomputing*, 469:28–51, 2022.
- [31] Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7765–7773, 2018.
- [32] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.
- [33] 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.
- [34] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.
- [35] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.
- [36] Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In *International Conference on Machine Learning*, pages 4548–4557. PMLR, 2018.
- [37] 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.
- [38] Vishal Thengane, Salman Khan, Munawar Hayat, and Fahad Khan. Clip model is an efficient continual learner. *arXiv preprint arXiv:2210.03114*, 2022.
- [39] Hugo Touvron, Matthieu Cord, Douze Matthijs, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *International Conference on Machine Learning*, 2020.
- [40] 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.
- [41] Liyuan Wang, Mingtian Zhang, Zhongfan Jia, Qian Li, Chenglong Bao, Kaisheng Ma, Jun Zhu, and Yi Zhong. Afec: Active forgetting of negative transfer in continual learning. *Advances in Neural Information Processing Systems*, 34:22379–22391, 2021.

- [42] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *arXiv preprint arXiv:2302.00487*, 2023.
- [43] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 139–149, 2022.
- [44] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2575–2584, 2020.
- [45] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 374–382, 2019.
- [46] Peng Xu, Xiatian Zhu, and David A Clifton. Multi-modal learning with transformers: A survey. *arXiv preprint arXiv:2206.06488*, 2022.
- [47] S YANG, P SUN, Y JIANG, X XIA, and P Luo. Objects in semantic topology. In *International Conference on Learning Representation (ICLR)*, 2022.
- [48] 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.
- [49] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International conference on machine learning*, pages 3987–3995. PMLR, 2017.
- [50] Xunlin Zhan, Yangxin Wu, Xiao Dong, Yunchao Wei, Minlong Lu, Yichi Zhang, Hang Xu, and Xiaodan Liang. Product1m: Towards weakly supervised instance-level product retrieval via cross-modal pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11782–11791, 2021.
- [51] Gengwei Zhang, Liyuan Wang, Guoliang Kang, Ling Chen, and Yunchao Wei. Slca: Slow learner with classifier alignment for continual learning on a pre-trained model. *arXiv preprint arXiv:2303.05118*, 2023.
- [52] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021.
- [53] Zekang Zhang, Guangyu Gao, Zhiyuan Fang, Jianbo Jiao, and Yunchao Wei. Mining unseen classes via regional objectness: A simple baseline for incremental segmentation. *Advances in Neural Information Processing Systems*, 35:24340–24353, 2022.
- [54] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11953–11962, 2022.
- [55] Da-Wei Zhou, Qi-Wei Wang, Zhi-Hong Qi, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Deep class-incremental learning: A survey. *arXiv preprint arXiv:2302.03648*, 2023.