MaskCLIP: Masked Self-Distillation Advances Contrastive Language-Image Pretraining

Xiaoyi Dong[1*†], Jianmin Bao[2*], Yinglin Zheng[3], Ting Zhang[2], Dongdong Chen[4,†], Hao Yang[2], Ming Zeng[3], Weiming Zhang[1], Lu Yuan[4], Dong Chen[2], Fang Wen[2], Nenghai Yu[1]
[1]University of Science and Technology of China [2]Microsoft Research Asia
[3]Xiamen University [4]Microsoft Cloud + AI
{dlight@mail., zhangwm@, ynh@}.ustc.edu.cn cddlyf@gmail.com
{jianbao, ting.zhang, luyuan, doch, fangwen}@microsoft.com
{zhengyinglin@stu., zengming@}xmu.edu.cn yanghao.alexis@foxmail.com

Abstract

This paper presents a simple yet effective framework MaskCLIP, which incorporates a newly proposed masked self-distillation into contrastive language-image pretraining. The core idea of masked self-distillation is to distill representation from a full image to the representation predicted from a masked image. Such incorporation enjoys two vital benefits. First, masked self-distillation targets local patch representation learning, which is complementary to vision-language contrastive focusing on text-related representation. Second, masked self-distillation is also consistent with vision-language contrastive from the perspective of training objective as both utilize the visual encoder for feature aligning, and thus is able to learn local semantics getting indirect supervision from the language. We provide specially designed experiments with a comprehensive analysis to validate the two benefits. Symmetrically, we also introduce the local semantic supervision into the text branch, which further improves the pretraining performance. With extensive experiments, we show that MaskCLIP, when applied to various challenging downstream tasks, achieves superior results in linear probing, finetuning, and zero-shot performance with the guidance of the language encoder. Code will be release at `https://github.com/LightDXY/MaskCLIP`.

1. Introduction

Vision-language (VL) contrastive learning [31, 51] has shown remarkable success in pretraining for various tasks. With large-scale image-text pairs available on the Internet, the model composed of a simple dual encoder design learns strong semantic prior by aligning between image and text. The resulting visual encoder not only exhibits excellent linear probing and finetuning performance, but also enables impressive zero-shot performance with the guidance of the language encoder, showing the generality of natural language and its ability to supervise a wide range of visual concepts.

Nonetheless, the associated language description, though providing richer information than mere class labels, still can hardly describe all the information in the corresponding image, as images are continuous signals with fine-grained details and complex semantics. As a result, the VL contrastive by aligning global representations may only focus on the text-described objects and ignore the rest which might be useful for downstream tasks.

In this paper, we are interested in how to fully leverage the image itself to facilitate the VL contrastive to further improve the transfer capability. (1) Firstly, the learned feature representation shall characterize local patches, serving as a complementary for global representation in VL contrastive. Inspired by the recent success of masked image modeling [4, 19, 26, 51, 60, 61] in learning patch representations, we also randomly mask the input image with a large portion to force the visual encoder to focus on the remaining visible patches. (2) Secondly, the learned representation for local patches shall possess semantic meanings, being consistent with the global representation receiving semantic text supervision. We bring mean teacher self-distillation [25, 57] to supervise the learned patch representations with the visual feature representations, enabling implicit supervision from natural language. The resulting objective is denoted as *masked self-distillation* where the student model and the teacher model come from the same neural networks and the knowledge is distilled from the full image (fed to the teacher model) to the masked image (fed to student model). To this end, we introduce MaskCLIP by incorporating masked self-

*Equal contribution, † Corresponding Author
†Work done during an internship at Microsoft Research Asia

distillation into VL contrastive to advance the transferable visual encoder.

There are several recent attempts [49, 68] also exploring the capability of the visual encoder under natural language supervision. The common approach is to introduce contrastive learning or masked image modeling on the vision side together with contrastive language-image pretraining. However, the performance indeed improves based on CLIP but does not as well as our masked self-distillation. We argue that (1) the contrastive learning objective based on central crop augmentation actually learns global representations for salient objects while lack of attention on the surrounding backgrounds [11]; and (2) masked image modeling usually needs to remap the learned representation to pixels [26] or discrete tokens [4]. Such low-level prediction target is inefficient for semantic feature learning and thus also conflicts with high-level language supervision in VL contrastive. A brief illustration is presented in Figure 1. In the experiments, we conduct comprehensive ablations to analyze the difference and provide numerical and visual evidence for better understanding.

Symmetrically, we argue that local semantic supervision on the text branch is also helpful for the text encoder and eventually beneficial for zero-shot performance. So we introduce the same mask-data-modeling format supervision into the text branch as well. Different from images where the pixel is low-level signal, the words crafted by human beings are already highly semantic, so we use the tokenized word piece as the prediction target directly, following the well-studied mask language modeling method BERT. Meanwhile, to reduce the output conflicts between contrastive learning and mask language modeling, we introduce a small decoder for the mask language modeling branch.

We train our MaskCLIP on a subset of a publicly available image-text pairs dataset, YFCC [58], and thoroughly evaluate the transfer ability of visual representations on several vision benchmarks: ImageNet-1K [17] for classification, ADE20K [69] for semantic segmentation, MS-COCO [40] for detection and segmentation, as well as a batch of other classification benchmarks. When it comes to ImageNet-1K [17] classification, MaskCLIP achieves $+6.9\%$, $+7.2\%$, $+1.3\%$ higher than CLIP for zero-shot transfer, linear probing, and finetuning respectively. For vision downstream tasks, we reach $+2.7$ mIoU on ADE20K [69] and $+1.8$ AP^b, $+1.4$ AP^m on MS-COCO [40]. For vision-language tasks, MaskCLIP achieves $+6.1\%$ average zero-shot accuracy on 20 datasets, and $+17.2\%$, $+12.8\%$ rank@1 improvement on the Flickr30K [67] image-test retrieval. In the recent Image Classification in the Wild challenge academic track, our MaskCLIP gets the 1_{st} result with 48.9% TOP-1 average accuracy, surpassing the second team with 3.4%.

In summary, the major contributions of this work are:

1. We present a novel vision-language pretraining framework MaskCLIP, by introducing masked self-distillation objective to facilitate VL contrastive for better transferable visual models.
2. We present extensive ablation studies on MaskCLIP variants and provide in-depth analysis numerically and visually to help understand how the proposed masked self-distillation assists VL contrastive.
3. We demonstrate our MaskCLIP on tens of benchmarks, showing the superiority under all three settings: zero-shot, linear probing, and finetuning.

2. Related Work

Vision-language pretraining Recent years have seen rapid progress made in vision-language pretraining [15, 18, 33, 35–39, 44–46, 50, 55, 56, 72]. Several multiple cross-modality loss functions have been proposed for the training objective, such as image-text matching [15, 37, 44, 56, 64], masked language modeling [15, 37, 44, 55, 56], masked image modeling [15, 44, 55, 56], contrastive loss [35, 38, 39]. These objects are often mixed with each other to form a compound objective. While a variety of approaches have been proposed, few works investigate the performance on visual representation learning for image classification. Recently, CLIP [51] and ALIGN [31] show that the image-text contrastive learning objective achieves promising performance for visual representation learning. There are many following works proposed to further improve the pretraining performance, DeCLIP [70], SLIP [49], COTS [43], ViCHA [54], CYCLIP [24] use additional uni/multi-modality supervision to improve the model capability, and PyramidCLIP [23], KLITE [53], IDEA [30] seek to external knowledge from pre-trained models or datasets as the additional guidance. FILIP [66] and LOUPE [34] introduce fine-grained alignment to the model. Focusing on this research direction, we analyze the desired properties of supervision which could be complementary to CLIP, and propose the masked self-distillation objective incorporated with the image-text contrastive loss to further improve pretraining performance for various visual understanding tasks.

Self-supervised learning Self-supervised visual representation learning has attracted increasing attention over the past few years. The objective of the self-supervised learning is mainly divided into two categories: contrastive and generative [41]. The contrastive methods, such as MOCO [12, 27], SimCLR [9, 10], BYOL [25], SimSiam [13], and DINO [6] measure the similar and dissimilar samples by contrastive loss. Their success heavily depends on the strong data augmentation. The generative methods, such as BEiT [4], MAE [26], PeCo [19], BEVT [60], BootMAE [20] and MaskFeat [61] leverage masked image modeling to reconstruct the remaining masked part of its original input from the given visible parts. The generative methods show more

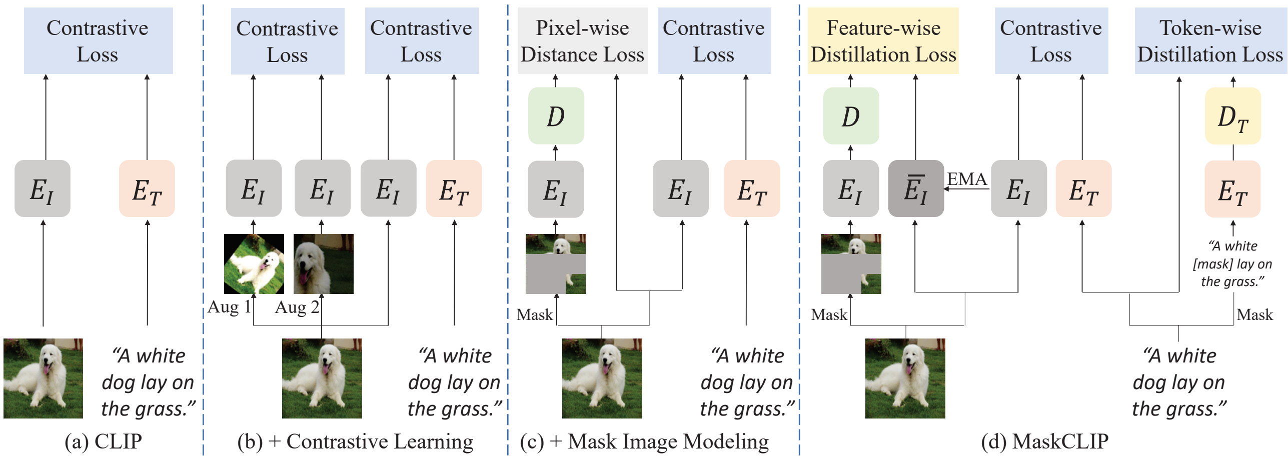

Figure 1. Pipeline comparison between combination CLIP with different vision self-supervised learning methods. (a) Vanilla CLIP. (b) CLIP + contrastive learning. (c) CLIP + pixel prediction mask image modeling. (d) CLIP + mask self-distillation, *i.e*. MaskCLIP. The E_T, E_I is the text encoder and image encoder respectively, and all the E_I, E_T within each pipeline share the weight. $\bar{E}_I$ is the mean-teacher model, whose weight is updated by the exponential moving average of E_I and does not require gradient.

promising transfer performance than the contrastive methods, as generative objective learns patch representations while contrastive objective focuses on learning centric global representations [11].

Self-knowledge distillation Self-knowledge distillation [32] aims to distill the knowledge in a model itself and uses it for training the model. Instead of distilling knowledge from a pretrained teacher model [29], self-knowledge distillation regards a temporal ensemble of the student model as the teacher. It means that a student model becomes a teacher model itself, which gradually utilizes its own knowledge for softening the hard targets to be more informative during training. Self-knowledge distillation has been explored in semi-supervised learning [57], contrastive learning [16, 35], self-supervised learning [3, 7]. In this paper, we use visual features supervised by natural language for guidance in masked self-distillation which naturally fit VL contrastive to learn more transferable visual representations.

3. MaskCLIP

We introduce MaskCLIP, a novel framework that learns visual representations. The core part of MaskCLIP is its backbone image encoder, denoted by E_I as shown in Figure 1. It obtains the transferable capability during pretraining that could benefit downstream vision tasks. Following recent self-supervised approaches [4, 14, 26, 49], we implement the backbone E_I as a Vision Transformer (ViT) [22]. The prediction results from E_I given an input image I then should be a collection of visual feature tokens, represented as

$$E_I(I) = \{f_{cls}, f_1, f_2, \dots, f_N\}. \quad (1)$$

Here *cls* is short for class token. $1, \dots, N$ are the indexes of the non-class tokens.

The rest of this section starts with the utilization of language supervision. More shall be emphasized on the masked self-distillation, which we deem crucial for visual pretraining.

3.1. Vision-language Contrastive

Following [31, 51], we introduce a Transformer-based text encoder E_T to leverage language knowledge. It aims to align the global feature representations of an image and a text with respect to some forms of similarity. Precisely, consider a given image-text pair $\{I, T\}$, besides extracting the visual feature representation $E_I(I)$ using the vision backbone as shown by Equation 1, we additionally use the text encoder E_T to extract linguistic features from the text T.

The mean feature of the two branches are regarded as the global representations and are fed into a projection head (implemented as a fully-connected layer) respectively to obtain the metric embeddings e^T and e^I. Image-text contrastive loss is employed to align them during pretraining. The loss can be formulated as $\mathcal{L}_T + \mathcal{L}_I$, with

$$\mathcal{L}_I = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(e_i^I e_i^T/\sigma)}{\sum_{j=1}^{B}\exp(e_i^I e_j^T/\sigma)}$$
$$\mathcal{L}_T = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(e_i^T e_i^I/\sigma)}{\sum_{j=1}^{B}\exp(e_i^T e_j^I/\sigma)}, \quad (2)$$

where B stands for the number of image-text pairs within a training mini-batch, i, j are indexes within the batch; σ stands for the temperature for the loss functions, which is learned together with all other parameters during training.

3.2. Masked Self-distillation for Visual Encoder

Knowledge distillation is a learning paradigm where a student model is trained to match the output of a given teacher model, so that the student model can be improved by the teacher. Instead of bringing in an external teacher, self-distillation methods such as [7, 25, 57] proposes using a *mean teacher* model that is derived from the student itself. In specific, the teacher shares the same structure with the student, while the parameters of the teacher are exponential moving averages (EMA) of the parameters from the student. In the following, we would use the term "EMA model" to represent such mean teacher model constructed from the student.

MaskCLIP leverages the mean teacher self-distillation to enhance its vision representations. Let $\bar{E}_I$ be the EMA model of the backbone encoder E_I. θ_t and $\bar{\theta}_t$ are the parameters of E_I and $\bar{E}_I$ at training step t. $\bar{\theta}_t$ is updated with

$$\bar{\theta}_t = \alpha\bar{\theta}_{t-1} + (1-\alpha)\theta_t, \tag{3}$$

where α is a hyper-parameter for smoothing updates. We propose to incorporate masked image modeling into self-distillation, resulting in *masked self-distillation* with asymmetric input for student model and teacher model.

In specific, considering a given input image I, we first feed it to the EMA model $\bar{E}_I$ (teacher model) to obtain the distillation targets. These target features can be represented as

$$\bar{E}_I(I) = \{\bar{f}_{cls}, \bar{f}_1, \bar{f}_2, \dots, \bar{f}_N\}. \tag{4}$$

In the meantime, we randomly mask a large portion of the input image patches and then feed it into the original backbone E_I (student model). Following [26], we only feed the visible (unmasked) patches, denoted by I', into the original backbone E_I to speed up computation and save memory. Let $\mathcal{M}$ be the indexes of all the masked tokens. These encoded features corresponding to visible tokens can then be denoted as $E_I(I') = \{f'_{cls}\} \bigcup \left\{f'_{k\notin\mathcal{M}}\right\}$. They are then joined with a shared and learnable feature vector, denoted as m, that represents mask tokens, to form a complete set of features $\{f'_{cls}, f'_1, f'_2, \dots, f'_N\}$, with $f'_{i\in\mathcal{M}} = m$. We attach positional embeddings onto all these tokens, and append a small Transformer D as a decoder to predict features of the masked region from the visible tokens, which could be formulated as

$$\begin{aligned}(D \circ E_I)(I') &= D\left(\{f'_{cls}, f'_1, f'_2, \dots, f'_N\}\right) \\ &= \{f''_{cls}, f''_1, f''_2, \dots, f''_N\}\,. \end{aligned} \tag{5}$$

Inspired by [71], we use an online quantizer $h()$ to transform the output features into a soft codewords distribution, and minimize the cross-entropy between the target features and the predicted features. Formally,

$$\mathcal{L}_{Dist} = \frac{1}{|\mathcal{M}|}\sum_{k\in\mathcal{M}} -\bar{h}(\bar{f}_k)^T log\, h(f''_k). \tag{6}$$

here the parameter of the teacher quantizer $\bar{h}()$ is also EMA updated by the online quantizer, similar to the teacher model.

3.3. Local Semantic Learning for Text Encoder

Besides the local semantic supervision for the visual encoder, we argue it is also helpful for the text encoder. So we introduce the BERT pretraining into the text branch. For the text $T = \{t_{sos}, t_1, t_2, ..., t_M, t_{eos}\}$, we denote the masked input as $T' = \{t'_{sos}, t'_1, t'_2, ..., t'_M, t'_{eos}\}$, where $t'_{i\in\mathcal{M}_T} = m_t$ and $t'_{i\notin\mathcal{M}_T} = t_i$, and $\mathcal{M}_T$ be the indexes of all the masked text tokens. The output feature of the encoder is $E_T(T')$.

To reduce the output conflict between the global image-text contrastive learning and the local mask language modeling, we further introduce a small text decoder, which shares the same architecture as the encoder but with only a few layers. So that the global prediction and local prediction are conducted at different layers. We denote the output feature as: $(D_T \circ E_T)(T') = \{t''_{sos}, t''_1, t''_2, ..., t''_M, t''_{eos}\}$ and the loss could be formulated as:

$$\mathcal{L}_{MLM} = \frac{1}{|\mathcal{M}_T|}\sum_{k\in\mathcal{M}_T} -t_k^T log\, t''_k. \tag{7}$$

3.4. Overall Loss Functions

Finally, we pretrain MaskCLIP with all these losses combined:

$$\mathcal{L}_I + \mathcal{L}_T + \lambda\mathcal{L}_{Dist} + \beta\mathcal{L}_{MLM}, \tag{8}$$

with λ, β being the hyper-parameter weighting between VL contrastive loss and self-supervised learning loss. All the components of MaskCLIP are trained from scratch, including the visual backbone E_I, the visual decoder D, the text encoder E_T, as well as the text decoder D_T.

4. Experiments

4.1. Setup

Model architecture. Our framework consists of the visual encoder E_I, the text encoder E_T, the visual decoder D, and the text decoder D_T. We adopt the widely used Transformer ViT-B/16 [22] for a fair comparison. It is composed of 12 layers, 768 width, and 12 head. The input image is 224×224 resolution and is further split into 14×14 patches with size 16×16. A learnable cls token is prepended to the 196 embeddings. For the text encoder, we adopt a 12-layer, 512-width, and 8-head Transformer following CLIP [51], and the text decoder has 4 layers. The number of text tokens is fixed to 77 with necessary truncations or paddings. For the image decoder, we directly use a one-layer Vision Transformer.

Pretraining details. We train our proposed MaskCLIP from scratch for 25 epochs, the batch size is fixed to 4096 for all the experiments. The masks used in the mask self-distillation

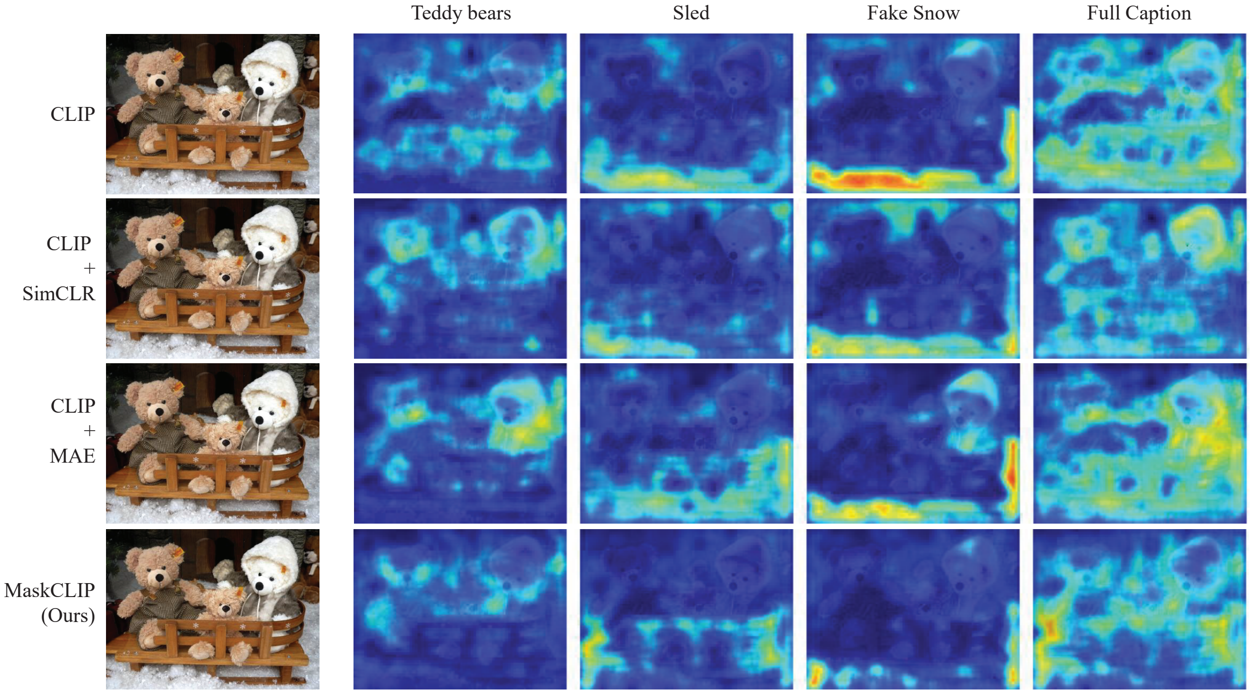

Figure 2. Visualization of the similarity between text and image features. The images and captions are from the MS-COCO val set. Here we show the image feature similarity with both full caption and different objects in it. The caption is "Three teddy bears sit in a sled in snow". More results could be found in the supplemental materials.

	Training		IN-1K			Flicker30K	
	Memory	Time	0-shot	Linear	Finetune	I2T	T2I
CLIP	14G	1.00×	37.6	66.5	82.3	52.9	32.8
CLIP+SimCLR	30G	2.67×	42.8	72.1	82.6	58.6	41.3
CLIP+MAE	16G	1.30×	42.1	68.5	83.2	57.3	41.1
MaskCLIP	19G	1.75×	**44.5**	**73.7**	**83.6**	**70.1**	**45.6**

Table 1. Results of boosting CLIP with different kinds of vision self-supervised learning methods.

branch and mask language modeling branch are random mask with a mask ratio of 75% and 20%. We pretrain all the models with the commonly used YFCC15M dataset, which is flited from the YFCC100M [58] dataset by [51].

Downstream details. We evaluated MaskCLIP on several downstream datasets, including ImageNet-1K [17], ADE20K [69], MS-COCO [40], Flicr30K [67] *et al*. For ImageNet-1K, we report zero-shot, linear probing, and finetuning performance. The zero-shot is conducted following the label prompt setting in SLIP [49]. For linear probing, we fix the backbone and train a new linear classifier for 90 epochs. For finetuning, we follow the setting in BEiT [4] and finetune the model for 100 epochs with a layer-decayed learning rate. See supplemental materials for more details.

4.2. Analysis

We first present our analysis by studying different ways of boosting CLIP. The baseline is CLIP [51] trained on the YFCC-15M. Besides the introduced masked self-distillation, we consider two other popular methods: (1) SimCLR [9], a representative contrastive method; and (2) MAE [26] the state-of-the-art masked image modeling approaches. All the compared methods are trained on the YFCC-15M for a fair comparison. We have the following observations.

Vision self-supervision helps VL contrastive. We evaluate the models on both vision task ImageNet-1K [17] classification and vision-language task image-text retrieval on Flicker30K [67] and present the comparison in Table 1. All the added vision self-supervision, regardless of contrastive or generative, improves the baseline CLIP. Among them, our proposed MaskCLIP achieves the best results in terms of all the evaluation metrics, outperforming CLIP with +6.9%, +7.2%, + 1.3% on ImageNet-1K classification for zero-shot, linear probing, and finetuning respectively, and +17.2%, +12.8% on Flicker30K for image-to-text retrieval and text-to-image retrieval. We also report the training GPU memory usage and time-consuming cost in Table 1. It is worth noting that the contrastive model (CLIP+SimCLR) compares two additional views of the input image, resulting in larger GPU memory usage and longer training time.

Masked image modeling is able to learn representations for local patches. We argue that the image encoder only pays attention to the text-described objects under VL contrastive due to sparse text description and to the centric objects under image contrastive due to central-crop augmenta-

Method	Objective	ADE20K mIoU	Pascal mIoU
CLIP	Global	7.2	13.5
CLIP + SimCLR	Global + Global	6.3	11.9
CLIP + MAE	Global + Pixel-wise Local	8.3	16.4
MaskCLIP	Global + Token-wise Local	10.2	17.2

Table 2. Annotation-free zero-shot segmentation results on ADE20K and Pascal Context.

tion. In contrast, masked image modeling forces the image encoder to focus on local patches using token-wise objectives by mandatorily masking a large portion of patches. Here, we provide numerical comparisons for evidence. We conduct an "Annotation-free zero-shot segmentation" experiment to test the zero-shot segmentation. The results on such a dense prediction task would better reveal the ability of local patch representations than global classification. Following the design in DenseCLIP [70], we use the prompted label feature as the linear classification weight to realize segmentation, without any training procedure. We evaluate the performance on two widely used datasets: ADE20K [69] and Pascal Context [48]. The results are shown in Table.2. We can see that equipped with masked image modeling, our MaskCLIP as well as CLIP+MAE achieves better results than CLIP and CLIP+SimCLR, validating our hypothesis.

Masked self-distillation learns semantic representations for local patches. Our masked self-distillation predicts visual features dynamically outputted by the visual encoder and thus implicitly gets supervision from the text side via VL contrastive. While MAE predicts fixed low-level pixels, making it inefficient to learn semantic representations (as the objective may force the representation to memorize low-level details) and thus causing conflict with VL contrastive. To show this, we select images from MS-COCO [40] and calculate the feature similarity between image features and their corresponding caption features. We also select objects in the caption, prompt it to a new caption, such as "a photo of teddy bears", and calculate the similarities. An example is shown in Figure 2 (More can be found in the supplementary material). Comparing MaskCLIP with CLIP+MAE in the fourth column, we can see that CLIP+MAE uses color as evidence and fails to distinguish the white teddy bear from the white snow. While our MaskCLIP successfully differentiates the two objects, suggesting ours learn more semantic features. On the other hand, the superior results of MaskCLIP shown in Table 1 and Table 2 also validate this. It is worth mentioning that CLIP and CLIP+SimCLR fail to have a correct response partition for different single objects like MaskCLIP, further justifying our second observation.

4.3. Comparison with Previous Methods

To show the effectiveness of MaskCLIP as a general vision-language pretrain method, we conduct experiments on both vision tasks and vision-language tasks. For vision

Method	Epoch	IN-1K 0-Shot	IN-1K Lin	IN-1K FT	ADE20K mIoU	MS-COCO AP^b	MS-COCO AP^m
DeiT [59]	300*	–	–	81.8	47.4	44.1	39.8
SimCLR [9]	25	–	64.0	82.5	48.0	44.6	40.2
MAE [26]	25	–	56.2	82.5	46.5	43.2	39.1
CLIP [51]	25	37.6	66.5	82.3	47.8	43.6	39.5
SLIP [49]	25	42.8	72.1	82.6	48.5	44.0	40.3
MaskCLIP	25	**44.5**	**73.7**	**83.6**	**50.5**	**45.4**	**40.9**

Table 3. Comparison with previous methods, including supervised baselines, self-supervised pretraining methods, and vision-language pretraining methods. * is the epoch of the supervised baseline on ImageNet-1K.

tasks, we report results on ImageNet-1K [17] classification, MS-COCO [40] object detection, and ADE20K [69] semantic segmentation. For vision-language tasks, we report zero-shot results on recent challenging ICinW 20 datasets benchmark and image-text retrieval results on Flickr30K [67] and MS-COCO [40]. In the following, we compare with the supervised baseline DeiT [59], self-supervised methods SimCLR [9] and MAE [26], and vision-language methods CLIP [51] and SLIP [49]. For a fair comparison, we train SimCLR and MAE on YFCC-15M [58] with the same epochs.

Classification on ImageNet-1K. As shown in Table 3, MaskCLIP benefits from the advantages of both VL pretraining and image mask self-distillation that shows strong performance on all the metrics. For zero-shot tasks, MaskCLIP outperforms CLIP by $+6.9\%$ with 25 epoch training and achieves $+1.7\%$ higher than the recent work SLIP. When it comes to finetune, MaskCLIP reaches 83.6% top-1 accuracy, and outperforms CLIP by $+1.3\%$.

Semantic segmentation on ADE20K. Then we apply our MaskCLIP to the semantic segmentation task. Here we use the UperNet [63] framework with 512×512 input and end-to-end training for 160K iterations. The evaluation metric is the mean Intersection of Union (mIoU) and we report single-scale evaluation results here. The results are given in Table 3. Our method achieves 50.5 mIoU, $+2.7$ mIoU than our baseline method CLIP, and $+2.0$ mIoU than SLIP. This verifies the effectiveness of our introduced incorporation.

Object detection and instance segmentation on MS-COCO. We further investigate our transfer performance on object detection and instance segmentation in Table.3. Here we use Mask-RCNN [28] framework with single-scale input and $1\times$ schedule (12 epochs). Our method achieves 45.4 box AP and 40.9 mask AP, $+1.8/1.4$ better than CLIP, and $+1.4/0.6$ better than SLIP.

Zero-shot on small datasets. We also report zero-shot performance on 20 small datasets under the ICinW setting (see the introduction below) in Table 4. We find that all the methods perform poorly on some datasets such as Aircraft(1% acc for random guessing, we omit the description in the

	Average	Caltech-101	CIFAR-10	CIFAR-100	Country211	DTD	EuroSAT	FER-2013	Aircraft	Food-101	GTSRB	Memes	KittiDis	MNIST	Flowers	Pets	PatchCam	SST2	RESISC45	Cars	Voc2007
Pretraining on YFCC-15M																					
CLIP	34.0	58.6	68.5	36.9	10.8	21.4	30.5	16.9	5.1	51.6	6.5	51.1	25.9	5.0	52.7	28.6	51.7	**52.5**	22.4	4.5	79.1
SLIP	37.8	70.9	**82.6**	48.6	11.8	26.6	19.8	18.1	5.6	59.9	**12.6**	51.8	29.4	**9.8**	56.3	31.4	**55.3**	51.5	28.5	5.4	80.5
MaskCLIP	**40.1**	**72.0**	80.2	**57.5**	**12.6**	**27.9**	**44.0**	**20.3**	**6.1**	**64.9**	8.5	**52.0**	**34.3**	4.9	**57.0**	**34.3**	50.1	49.9	**35.7**	**6.7**	**82.1**
Pretraining on ICinW Academic Track Stting: YFCC-15M , GCC3M+12M, ImageNet-21K(ImageNet-1K is removed)																					
1st MaskCLIP	**48.9**	86.4	**95.3**	**78.3**	11.6	**33.0**	**57.7**	18.8	**8.0**	**78.9**	17.3	**52.8**	16.0	7.3	74.2	**74.4**	**52.1**	46.2	**54.3**	**26.5**	**82.3**
2nd KLITE*	45.5	**87.4**	92.7	68.8	8.2	32.2	27.9	17.4	4.3	72.4	11.4	48.4	**31.1**	12.8	**75.6**	65.9	50.6	**52.9**	44.4	10.2	**82.3**
3rd YT-CLIP	44.5	77.8	83.5	58.4	**11.9**	31.9	40.7	27.1	6.9	68.7	**18.8**	52.3	9.1	**18.8**	53.1	69.3	51.5	50.3	52.7	19.7	79.3
4th UniCL†	44.0	84.8	90.2	67.8	6.7	25.4	35.3	**30.8**	3.5	68.3	11.1	51.0	17.9	11.3	71.7	44.9	52.1	49.5	41.4	24.2	81.3
5th Gramer*	43.2	83.9	92.9	69.5	7.3	25.5	24.4	30.4	2.7	71.0	9.0	52.6	12.4	10.1	70.4	52.4	50.6	50.1	44.8	13.8	81.3

Table 4. Zero-shot evaluation on ICinW classification benchmarks. Best results in **bold**. * indicates using Swin-B as the backbone, † indicates using Focal-B as the backbone.

	Training	Flickr30K						MS-COCO					
		Image-to-text			Text-to-image			Image-to-text			Text-to-image		
	Epoch	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP [51]	25	52.9	79.6	87.2	32.8	60.8	71.2	27.5	53.5	65.0	17.7	38.8	50.5
SLIP [49]	25	58.6	85.1	91.7	41.3	68.7	78.6	33.4	59.8	70.6	21.5	44.4	56.3
MaskCLIP	25	**70.1**	**90.3**	**95.3**	**45.6**	**73.4**	**82.1**	**41.4**	**67.9**	**77.5**	**25.5**	**49.7**	**61.3**

Table 5. Results of zero-shot image-text retrieval on Flickr30K and MS-COCO datasets. Best results in **bold**.

following), Fer(24.7%), Country211(0.5%), GTSRB(5.9%), Cars(0.8%). This might be caused by the data domain gap that the YFCC-15M contains few related images and descriptions. For the rest of the datasets, all the methods get reasonable performance and our MaskCLIP gets the best performance on most datasets.

Image Classification in the Wild (ICinW) Challenge The ICinW challenge [1] is a newly proposed visual pretraining benchmark, which contains 20 diverse downstream classification datasets, measuring the ability of pre-training models on both the prediction accuracy and their transfer efficiency in a new task. The pretraining is limited to three datasets: YFCC-15M [58], GCC3M [52]+12M [8] and ImageNet-21K [17] (ImageNet-1K data is excluded). We pretrain our MaskCLIP on it and get the $\mathbf{1}_{st}$ result in the zero-shot track [2] (we submit the results anonymously). As shown in Table 4, the 2_{nd} team KLITE uses a strong Swin-B [42] as the backbone and additional knowledge from GPT-3 [5] and Wiktionary [47], and the 4_{th} use the strong Focal-B [65] as the backbone, while our MaskCLIP greatly outperforms these methods with a simple ViT-B backbone and no additional knowledge.

Zero-shot on text-image retrieval. We further report the zero-shot text-image retrieval results on two benchmark datasets, Flicr30K [67] and MS-COCO [40]. We find that the text without any prefixes or suffixes works well for all the models. Table 5 shows the results. We can see that MaskCLIP exhibits a strong zero-shot performance. For example, with 25 epochs training, MaskCLIP reaches 41.4% Rank@1 image-to-text accuracy on MS-COCO, outperforming CLIP with 13.9%, and 25.5% Rank@1 text-to-image accuracy, +7.8% higher than CLIP.

4.4. Ablations

We compare our default settings with other alternatives to justify the efficacy of our model designs.

Training objectives ablation. As shown in Table.6a, when we remove the mask language modeling loss $\mathcal{L}_{\text{MLM}}$, the performance of the image-text task drops, including the zero-shot accuracy and retrial performance. While benefiting from the distillation loss, the finetuning performance on ImageNet-1K is not influenced. When we remove the distillation loss $\mathcal{L}_{\text{Dis}}$, we observe a performance drop on all tasks, especially the finetuning results.

Distillation loss format. Different from previous methods [3, 21, 26] that calculate the per-element distance as the loss function, we use an online tokenizer to map the feature to soft codewords and use the cross-entropy loss as the supervision. Here we study their difference in Table.6b. We find that although they get similar fine-tuning performance, the CE loss gets better zero-shot and linear probing performance. The reason may be that the per-element MSE loss leads the model to fit some unnecessary details of the target feature, while the CE loss with soft tokenizer helps the model to focus more on the important feature.

Distillation & MLM loss weight. Here we set the loss weight of the CLIP branch as 1 and study the loss weight of the two additional branches. As shown in Table.6e and Table.6f, setting $\lambda = 1$ or $\beta = 1$ emphasize too much on new tasks, which mislead the model to a wrong converge

Model	0-Shot	FT	I2T/T2I
MaskCLIP	44.5	83.6	70.1/45.6
w/o $\mathcal{L}_{\text{MLM}}$	42.8	83.6	65.0/41.6
w/o $\mathcal{L}_{\text{Dis}}$	42.0	82.4	65.4/40.5

(a) **Training Objectives ablation**. Both is necessary for MaskCLIP.

Loss	0-Shot	Lin	FT
MSE	43.8	73.2	83.6
CE	44.5	73.7	83.6

(b) **Distillation loss format**. The online tokenizer with cross-entropy loss works slightly better than MSE loss.

Depth	0-Shot	Lin	FT
1	44.5	73.7	83.6
2	43.7	72.9	83.4
4	43.5	72.5	83.3

(c) **Visual decoder Depth**. A shallow decoder gets better performance.

Depth	0-Shot	I2T/T2I
0	43.5	65.2/44.1
1	44.3	70.4/45.3
2	44.3	70.2/45.4
4	44.5	70.1/45.6
8	44.2	67.5/44.7

(d) **Text decoder depth**. The decoder is necessary and a shallow one works better.

Weight	0-Shot	Lin	FT
1	38.5	68.2	82.5
0.1	44.4	73.5	83.5
0.05	44.5	73.7	83.6
0.01	43.6	73.0	83.4

(e) **Distillation loss weight**. A small loss weight works well for MaskCLIP.

Weight	0-Shot	I2T/T2I
1	36.5	51.7/32.1
0.1	44.3	69.2/45.9
0.05	44.5	70.1/45.6
0.01	43.2	70.6/45.6

(f) **MLM loss weight**. A small loss weight works better.

Table 6. **MaskCLIP ablation experiments** with YFCC-15M dataset. We report zero-shot(0-Shot), fine-tuning (FT), and linear probing (Lin) accuracy (%) for image-encoder-related ablation. And zero shot image-to-text, text-to-image retrieval (I2T/T2I) for text encoder-related ablations. Default settings are marked in gray .

direction, resulting in poor performance. When we reduce the loss weight by $10\times$, the two additional tasks are helpful for the model and show a consistent gain on all the metrics. We suspect this is because the CLIP loss requires two different capabilities: understanding the input content and aligning them into a shared feature space. And the goal of the two additional self-supervised learning tasks is to facilitate understanding.

Image & Text decoder depth. Then we study the influence of the decoder depth for both image and text decoders. As shown in Table.6c, we find the image decoder with only one layer works well, increasing the decoder depth leads to worse performance on all metrics. Similarly, Table.6d shows that the text branch benefits from a shallow decoder design. We argue that a too-deep decoder would make the encoder lazy, relying on the strong decoder to resolve the challenging mask feature/language modeling tasks. And the different depth choice between the image and text branches is caused by the framework difference: the image branch sees the mask tokens at the decoder, while the text branch takes the mask tokens as the encoder input. Note that if we remove the text decoder, the performance gets worse. We think this is largely caused by the output conflict that the global recognition feature aggregation and local word prediction are conducted at the same layer.

Single-Stage v.s. two-Stage. Our MaskCLIP learns the VL contrastive and masked self-distillation simultaneously and jointly in a single stage. One possible variant is to first train CLIP and then use CLIP feature from the first stage to train masked image modeling as in [61, 62]. We report results on three datasets in Table 7. We can see that the second stage achieves better finetuning results compared with results from the stage one, showing the effectiveness of masked image modeling. Nonetheless, such two-stage training requires longer training time and loses the transfer capability in a zero-shot setting. In contrast, our MaskCLIP achieves superior results under all settings with fewer epochs.

Method		Epoch	IN-1K 0-shot	IN-1K FT	Flicker30K I2T	Flicker30K T2I	ADE20K 0-shot	ADE20K FT
Two-Stage	Stage1	25	37.6	82.3	52.9	32.8	7.2	47.8
	Stage2	25	—	83.4	—	—	—	48.2
MaskCLIP		25	**44.5**	**83.6**	**70.1**	**45.6**	**10.2**	**50.5**

Table 7. Comparison between two-stage method and our single-stage MaskCLIP.

5. Conclusion

We present MaskCLIP, a new VL pretraining framework that incorporates masked self-distillation into VL contrastive. We point out that masked self-distillation learns local semantics, fitting nicely to the VL contrastive that aims to learn global semantics, and this is supported with comprehensively designed experiments. We also utilize mask language modeling to enhance the text encoder which is critical for zero-shot performance. The resulting visual encoder shows strong transfer capability across widely adopted benchmarks for linear probing, fine-tuning, and also zero-shot evaluation.

Acknowledgements

This work was supported in part by the Natural Science Foundation of China under Grant U20B2047, 62072421, 62002334, 62102386, and 62121002, Key Research and Development program of Anhui Province under Grant 2022k07020008, Ant Group through CCF-Ant Innovative Research Program CCF-AFSG RF20210025, Alibaba Group through Alibaba Innovative Research Program. This work was also partly supported by Shenzhen Key Laboratory of Media Security, Shenzhen University, Shenzhen 518060, China.

References

[1] https://eval.ai/web/challenges/challenge-page/1832/overview.

[2] https://eval.ai/web/challenges/challenge-page/1832/leaderboard/4298.

[3] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. *arXiv preprint arXiv:2202.03555*, 2022.

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

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[8] 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.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[10] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020.

[11] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *arXiv preprint arXiv:2202.03026*, 2022.

[12] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[13] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning, 2020.

[14] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021.

[15] 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 *European conference on computer vision*, pages 104–120. Springer, 2020.

[16] Ruizhe Cheng, Bichen Wu, Peizhao Zhang, Peter Vajda, and Joseph E Gonzalez. Data-efficient language-supervised zero-shot learning with self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3119–3124, 2021.

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

[18] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11162–11173, 2021.

[19] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Peco: Perceptual codebook for bert pre-training of vision transformers. *arXiv preprint arXiv:2111.12710*, 2021.

[20] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Bootstrapped masked autoencoders for vision bert pretraining. *arXiv preprint arXiv:2207.07116*, 2022.

[21] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Bootstrapped masked autoencoders for vision bert pretraining. *arXiv preprint arXiv:2207.07116*, 2022.

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[23] Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, and Chunhua Shen. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *arXiv preprint arXiv:2204.14095*, 2022.

[24] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan A Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. *arXiv preprint arXiv:2205.14459*, 2022.

[25] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

[26] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.

[27] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

[28] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[29] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

[30] Xinyu Huang, Youcai Zhang, Ying Cheng, Weiwei Tian, Ruiwei Zhao, Rui Feng, Yuejie Zhang, Yaqian Li, Yandong Guo, and Xiaobo Zhang. Idea: Increasing text diversity via online multi-label recognition for vision-language pre-training. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4573–4583.

[31] 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.

[32] Kyungyul Kim, ByeongMoon Ji, Doyoung Yoon, and Sangheum Hwang. Self-knowledge distillation: A simple way for better generalization. *arXiv preprint arXiv:2006.12000*, 3, 2020.

[33] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344, 2020.

[34] Juncheng Li, Xin He, Longhui Wei, Long Qian, Linchao Zhu, Lingxi Xie, Yueting Zhuang, Qi Tian, and Siliang Tang. Fine-grained semantically aligned vision-language pre-training. *arXiv preprint arXiv:2208.02515*, 2022.

[35] 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. *Advances in Neural Information Processing Systems*, 34, 2021.

[36] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020.

[37] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[38] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*, 2020.

[39] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

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

[41] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[42] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[43] Haoyu Lu, Nanyi Fei, Yuqi Huo, Yizhao Gao, Zhiwu Lu, and Ji-Rong Wen. Cots: Collaborative two-stream vision-language pre-training model for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15692–15701, 2022.

[44] 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.

[45] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446, 2020.

[46] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020.

[47] Christian M Meyer and Iryna Gurevych. *Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography*. na, 2012.

[48] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 891–898, 2014.

[49] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. *arXiv preprint arXiv:2112.12750*, 2021.

[50] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*, 2020.

[51] 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.

[52] 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.

[53] Sheng Shen, Chunyuan Li, Xiaowei Hu, Yujia Xie, Jianwei Yang, Pengchuan Zhang, Anna Rohrbach, Zhe Gan, Lijuan Wang, Lu Yuan, et al. K-lite: Learning transfer-

able visual models with external knowledge. *arXiv preprint arXiv:2204.09222*, 2022.

[54] Mustafa Shukor, Guillaume Couairon, and Matthieu Cord. Efficient vision-language pretraining with visual concepts and hierarchical alignment. *arXiv preprint arXiv:2208.13628*, 2022.

[55] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

[56] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[57] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[58] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[59] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[60] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. Bevt: Bert pretraining of video transformers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2022)*, 2022.

[61] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. *arXiv preprint arXiv:2112.09133*, 2021.

[62] Longhui Wei, Lingxi Xie, Wengang Zhou, Houqiang Li, and Qi Tian. Mvp: Multimodality-guided visual pre-training. *arXiv preprint arXiv:2203.05175*, 2022.

[63] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018.

[64] Haiyang Xu, Ming Yan, Chenliang Li, Bin Bi, Songfang Huang, Wenming Xiao, and Fei Huang. E2e-vlp: End-to-end vision-language pre-training enhanced by visual learning. *arXiv preprint arXiv:2106.01804*, 2021.

[65] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal self-attention for local-global interactions in vision transformers. *arXiv preprint arXiv:2107.00641*, 2021.

[66] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.

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

[68] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. *arXiv preprint arXiv:2112.03109*, 2021.

[69] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.

[70] Chong Zhou, Chen Change Loy, and Bo Dai. Denseclip: Extract free dense labels from clip. *arXiv preprint arXiv:2112.01071*, 2021.

[71] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image bert pre-training with online tokenizer. In *International Conference on Learning Representations*, 2021.

[72] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049, 2020.