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

Xiaoyi Dong^{1*}, Yinglin Zheng^{2*}, Jianmin Bao³, Ting Zhang³, Dongdong Chen⁴, Hao Yang³,
Ming Zeng², Weiming Zhang¹, Lu Yuan⁴, Dong Chen³, Fang Wen³, Nenghai Yu¹

¹University of Science and Technology of China ²Xiamen University

³Microsoft Research Asia ⁴Microsoft 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

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. Empirically, we show that MaskCLIP, when applied to various challenging downstream tasks, achieves superior results in linear probing, finetuning as well as the zero-shot performance with the guidance of the language encoder.

1 Introduction

Vision-language (VL) contrastive learning [24, 41] 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 [2, 17, 20, 41, 47, 48] in learning patch representations, we also randomly mask the input image with a large portion to force the visual

*Work done during an internship at Microsoft Research Asia.

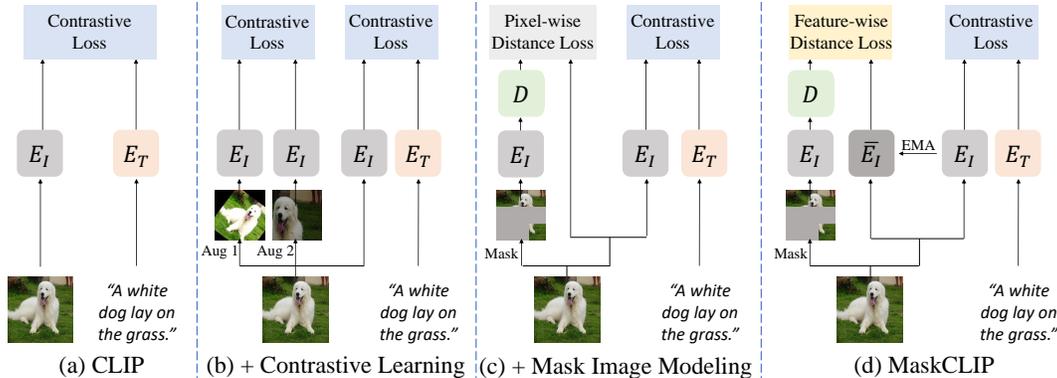


Figure 1: Pipeline comparison between combination CLIP with different vision self-supervised learning methods. (a) Vanilla CLIP. (b) CLIP + contrastive learning. (c) CLIP + pixel prediction mask image modeling. (d) CLIP + mask self-distillation, *i.e.*, MaskCLIP. The E_T , E_I is the text encoder and image encoder respectively, and all the E_I 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.

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 [19, 45] 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-distillation into VL contrastive to advance the transferable visual encoder.

There are several recent attempts [39, 54] 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 [8]; and (2) masked image modeling usually needs to remap the learned representation to pixels [20] or discrete tokens [2]. 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.

We train our MaskCLIP on a subset of a publicly available image-text pairs dataset, LAION [42], and thoroughly evaluate the transfer ability of visual representations on several vision benchmarks: ImageNet-1K [15] for classification, ADE20K [55] for semantic segmentation, MS-COCO [32] for detection and segmentation, as well as a batch of other classification benchmarks. When it comes to ImageNet-1K [15] classification, MaskCLIP achieves +5.8%, +4.3%, +1.2% percentage points higher than CLIP with respect to zero-shot transfer, linear probing, and finetuning. For vision downstream tasks, we reach +3.8 mIoU on ADE20K [55] and +2.0 AP^b, +1.4 AP^m on MS-COCO [32]. For vision-language tasks, MaskCLIP achieves +2.7% average zero-shot accuracy on 25 datasets, and +8.7%, +8.0% rank@1 improvement on the Flickr30K [53] image-test retrieval.

In summary, the major contributions of this work are:

1. We present a novel vision-language pretraining framework, MaskCLIP by introducing a 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 [12, 16, 26, 27, 28, 29, 30, 31, 35, 36, 37, 40, 43, 44, 57]. Several multiple cross-modality loss functions have been proposed for the training objective, such as image-text matching [12, 29, 35, 44, 51], masked language modeling [12, 29, 35, 43, 44], masked image modeling [12, 35, 43, 44], contrastive loss [27, 30, 31]. 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 [41], ALIGN [24] and SLIP [39] show that the image-text contrastive learning objective achieves promising performance for visual representation learning. Focusing on this research direction, we 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 [33]. The contrastive methods, such as MOCO [9, 21], SimCLR [6, 7], BYOL [19], SimSiam [10], and DINO [3] measure the similar and dissimilar samples by contrastive loss. Their success heavily depends on the strong data augmentation. The generative methods, such as BEiT [2], MAE [20], PeCo [17], BEVT [47] and MaskFeat [48] leverage masked image modeling to reconstruct the remaining masked part of its original input from the given visible parts. The generative methods show more promising transfer performance than the contrastive methods, as generative objective learns patch representations while contrastive objective focuses on learning centric global representations [8].

Self-knowledge distillation Self-knowledge distillation [25] 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 [23], 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 [45], contrastive learning [13, 27], self-supervised learning [1, 4]. 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 [2, 11, 20, 39], we implement the backbone E_I as a Vision Transformer (ViT) [18]. 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 [24, 41], 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 . We represent these linguistic features as $E_T(T) = \{f_1^T, f_2^T, \dots, f_M^T, f_{eos}^T\}$, where eos means the end of sequence token.

f_{eos}^T from the linguistic features and mean feature of the visual features 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

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

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

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 [4, 19, 45] 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

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

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, consider 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

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

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 [20], 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 $E_I(I') = \{f'_{cls}\} \cup \{f'_{k \notin \mathcal{M}}\}$. 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 one-layer Transformer D as a decoder to predict features of the masked region from the visible tokens, which could be formulated as

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

A distillation loss \mathcal{L}_{Dist} is imposed to make the predicted features from $(D \circ E_I)(I')$ match the target features generated by $\bar{E}_I(I)$ (Equation 4):

$$\mathcal{L}_{Dist} = \frac{1}{|\mathcal{M}|} \sum_{k \in \mathcal{M}} \text{SmoothL1}(f''_k, \text{StopGradient}(\bar{f}_k), \beta). \quad (6)$$

We utilize smooth L1 loss for feature matching and β is the smooth factor, we set $\beta = 2$ by default.

3.3 Overall Loss Functions

Finally, we pretrain MaskCLIP with all these losses combined:

$$\mathcal{L}_I + \mathcal{L}_T + \lambda \mathcal{L}_{Dist} \quad (7)$$

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

4 Experiments

4.1 Setup

Model architecture. Our framework consists of three modules, the visual encoder E_I , the text encoder E_T , and a visual decoder D . We adopt the widely used Transformer ViT-B/16 [18] 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 image patches with patch size 16×16 . A learnable cls token is prepended to the 196 embeddings. For the text encoder, we adopt the 12-layer, 512-width, and 8-head Transformer following CLIP [41]. The number of input text tokens is fixed to 77 with necessary truncations or paddings. For the decoder, we directly use a one-layer Vision Transformer for both simplicity and performance considerations.

Pretraining details. We train our proposed MaskCLIP from scratch and training for 25 epochs, the batch size is fixed to 4096 for all the experiments. The masks used in the mask self-distillation branch are random mask with a mask ratio of 75%. We pretrain all the models with a randomly sampled 20M subset from the recent image-text dataset LAION-400M [42]. We denote this subset as LAION-20M.

Downstream details. We evaluated MaskCLIP on several downstream datasets, including ImageNet-1K [15], ADE20K [55], MS-COCO [32], Flickr30K [53] *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 [39]. For linear probing, we fix the backbone and train a new linear classifier for 90 epochs. For finetuning, we follow the setting in BEiT [2] which finetune the model for 100 epochs with a layer decayed learning rate. The setting of the rest dataset is described in the corresponding section and please see the supplemental materials for more details.

4.2 Analysis

We first present our analysis by studying different ways of boosting CLIP. The baseline is CLIP [41] trained on LAION-20M. Besides the introduced masked self-distillation, we consider three other popular methods: (1) SimCLR [6], a representative contrastive method; and (2) MAE [20] and BEiT [2], two state-of-the-art masked image modeling approaches. All the compared methods are trained on LAION-20M 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 [15] classification and vision-language task image-text retrieval on Flickr30K [53] and present the comparison in Table 1. We can see that 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 +5.8%, +4.3%, +1.2% on ImageNet-1K classification for zero-shot, linear probing and finetuning respectively, and +8.7%, +8.0% on Flickr30K 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 augmentation. In contrast, masked image modeling forces the image encoder to focus on local patches using token-wise objective by mandatorily masking a large portion of patches. Here, we provide numerical comparison 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

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

	Training		IN-1K			Flicker30K	
	Memory	Time	0-shot	Linear	Finetune	I2T	T2I
Baseline (CLIP) [41]	14G	1.00×	40.8	68.2	82.6	56.2	40.1
+ Contrastive learning (CLIP+SimCLR [6])	30G	2.67×	43.1	70.4	83.2	61.1	45.5
+ Raw Pixel Prediction (CLIP+MAE [20])	16G	1.30×	43.5	69.6	83.5	59.7	44.0
+ Discrete Tokens Prediction (CLIP+BEiT [2])	24G	1.82×	42.7	69.5	83.2	57.8	43.4
+ Online Feature Prediction (MaskCLIP)	17G	1.56×	46.6	72.5	83.8	64.9	48.1

Table 2: Annotation-free zero-shot segmentation results on MS-COCO, ADE20K and Pascal Context.

Method	Objective	MS-COCO mIoU	ADE20K mIoU	Pascal Context mIoU
CLIP	Global	8.2	7.7	13.5
CLIP + SimCLR	Global + Global	8.8	6.8	12.3
CLIP + MAE	Global + Pixel-wise Local	11.8	8.6	16.8
MaskCLIP (Ours)	Global + Token-wise Local	11.8	11.2	17.7

patch representations than global classification. Following the design in DenseCLIP [56], we use the prompted label feature as the linear classification weight to realize segmentation, without any training procedure. We evaluate the performance on three widely used datasets: MS-COCO [32], ADE20K [55] and Pascal Context [38]. 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 introduced 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 [32] and calculate the feature similarity between image features and its 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 6 (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 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 Ablations

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 [48, 49]. We report results on three datasets in Table 3. 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 zero-shot setting. In contrast, our MaskCLIP achieves superior results under all settings with fewer epochs.

Data scaling and training time scaling. We study the scaling performance of MaskCLIP with respect to the training data size and also the training epochs. For training data, we random sample 10M, 20M and 50M from the LAION-400M dataset and train both CLIP and our MaskCLIP for 25 epochs. For training epochs, we set the number of epochs to 25, 50, 100 and train the model on LAION-20M. The results are shown in Figure 3. We can see that MaskCLIP gets consistent

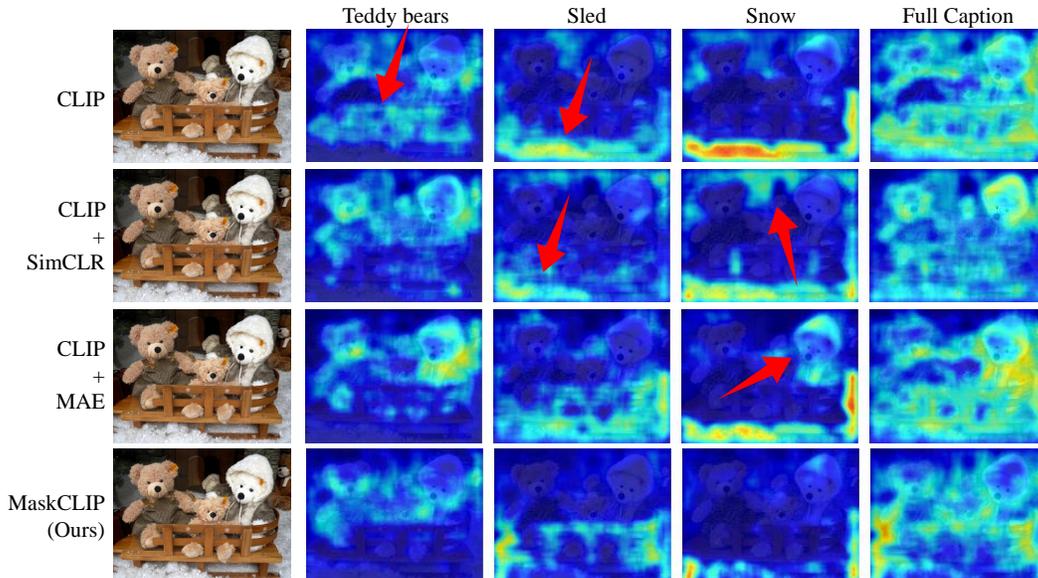


Figure 2: Visualization of the similarity between text and image features. The images and captions are from the MS-COCO val set. Here we show the image feature similarity with both full caption and different objects in it. The caption is “Three teddy bears sit in a sled in snow”. We use red arrows to point out the incorrect region. More results could be found in the supplemental materials.

Table 3: Comparison between two-stage method and our single-stage MaskCLIP.

Method		Epoch	IN-1K			Flicker30K		ADE20K	
			0-shot	Linear	Finetune	I2T	T2I	0-shot	Finetune
Two-Stage	Stage1	25	40.8	68.2	82.6	56.2	40.1	7.7	45.8
	Stage2	25	—	65.9	83.5	—	—	—	48.2
MaskCLIP(Single-Stage)		25	46.6	72.5	83.8	64.9	48.1	11.2	49.6

improvement over CLIP with either more data or more epochs. Furthermore, MaskCLIP is efficient that MaskCLIP with only 25 epochs performs better than CLIP trained for 100 epochs in all the cases. When it comes to data size, MaskCLIP with only 20M data shows comparable performance with CLIP with 50M data on zero-shot tasks, and outperforms it on the finetuning tasks.

4.4 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 tasks, we report results on ImageNet-1K [15] classification, MS-COCO [32] object detection, and ADE20K [55] semantic segmentation. For vision-language tasks, we report zero-shot results on 25 datasets and image-text retrieval results on Flickr30K [53] and MS-COCO [32]. In the following, we compare with the supervised baseline DeiT [46], self-supervised methods SimCLR [6] and MAE [20], and vision-language methods CLIP [41] and SLIP [39]. For fair comparison, we train SimCLR and MAE on LAION-20M with the same epochs.

Classification on ImageNet-1K. As shown in Table 4, 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 task, MaskCLIP outperforms CLIP with +5.8% with 25 epoch training, and +3.5% better than the concurrent work SLIP. When it comes to finetune, MaskCLIP reaches 84.1% top-1 accuracy with 100 epoch training, outperforms CLIP with +1.4%.

Semantic segmentation on ADE20K. Then we apply our MaskCLIP on semantic segmentation task. Here we use the UperNet [50] framework with 512×512 input and end-to-end training for 160K iterations. The evaluation metric is mean Intersection of Union (mIoU) and we report single-scale evaluation results here. The results are given in Table 4. Our method achieve 49.6 mIoU, +2.2 mIoU

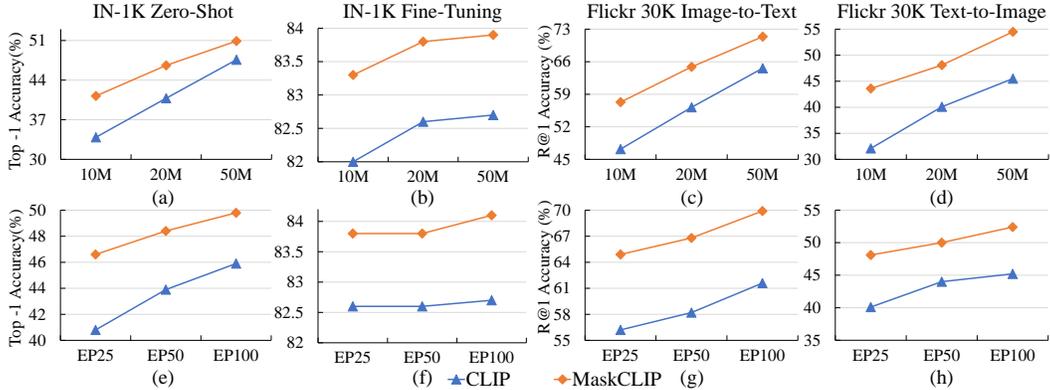


Figure 3: (a-d) Training data size scaling comparison on CLIP and MaskCLIP. (e-h) Training epoch scaling comparison. Here we report zero-shot and finetuning accuracy on ImageNet-1K and image-text retrieval accuracy on Flickr30K.

Table 4: Comparison with previous methods, including supervised baselines, self-supervised pre-training methods and vision-language pretraining methods. The *Epoch* column shows the epochs pretrained on the corresponding dataset and their equivalent epoch number on ImageNet-1K.

Method	Dataset	Epoch	0-Shot	IN-1K		ADE20K mIoU	MS-COCO	
				Linear	Finetune		AP ^b	AP ^m
DeiT [46]	IN1K-1.3M	300(300)	–	–	81.8	47.4	44.1	39.8
SimCLR [6]	IN1K-1.3M	300(300)	–	74.5	82.8	47.1	44.5	40.1
MAE [20]	IN1K-1.3M	1600(1600)	–	68.0	83.6	48.1	47.2	42.0
SimCLR [6]	LAION-20M	25(400)	–	58.2	82.2	45.7	43.5	39.5
MAE [20]	LAION-20M	25(400)	–	50.1	82.7	44.5	43.4	39.2
CLIP [41]	LAION-20M	25(400)	40.8	68.2	82.6	45.8	44.0	39.8
SLIP [39]	LAION-20M	25(400)	43.1	70.4	83.2	48.2	44.7	41.0
MaskCLIP	LAION-20M	25(400)	46.6	72.5	83.8	49.6	46.0	41.2
MAE [20]	LAION-20M	100(1600)	–	50.4	83.0	45.8	44.7	40.1
CLIP [41]	LAION-20M	100(1600)	45.9	68.9	82.7	46.9	45.2	40.4
MaskCLIP	LAION-20M	100(1600)	49.8	72.9	84.1	50.8	46.6	41.7

than supervised based methods. It is also +3.8 mIoU than our baseline method CLIP, and +1.4 mIoU than SLIP. When we extend the pretraining to 100 epochs, our MaskCLIP reaches a strong results of 50.8 mIoU, consistently better than CLIP with +3.9 mIoU. 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.4. Here we use Mask-RCNN [22] framework with single-scale input and $1\times$ schedule (12 epochs). The evaluation metric is box AP for detection and mask AP for segmentation. We find that our MaskCLIP trained on LAION-20M performs slightly worse than MAE trained on ImageNet-1K and this comparison has additional variable that the training dataset is different, we argue this may be caused by the dataset domain gap. For fair comparison with all models trained on LAION-20M, our MaskCLIP performs the best.

Zero-shot on small datasets. We also report zero-shot performance on 25 datasets in Table 5, following the setting in [39] that prompts the label of each dataset with a different context. 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 following), Fer(24.7%), Flowers(1.5%), Country211(0.5%), PCAM(50%). This might be caused by the data domain gap that the subset of LAION-400M we used does not contain 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.

Zero-shot on text-image retrieval. We further report the zero-shot text-image retrieval results on two benchmark datasets, Flickr30K [53] and MS-COCO [32]. We find that the text without any prefix or suffix works well for all the models. Table 6 shows the results. We can see that MaskCLIP exhibits

Table 5: Zero-shot evaluation on a variety of classification benchmarks. * indicates 100 epochs training results. Best results in **bold**. MaskCLIP outperforms CLIP and SLIP in most tasks, frequently with a significant margin.

	Food-101	CIFAR-10	CIFAR-100	CUB	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	MINIST	FER-2013	STL-10	EuroSAT	RESISC45	GTSRB	KITTI	Country211	PCAM	UCF101	Kinetics700	CLEVR	SST2	ImageNet	Average
CLIP [41]	59.7	82.9	56.9	26.4	50.7	57.8	4.9	33.8	62.2	77.7	5.2	35.3	29.2	90.9	44.3	42.0	28.7	36.5	7.2	57.7	46.1	29.8	13.7	49.4	40.8	42.7
SLIP [39]	61.5	88.2	62.4	21.7	53.3	57.1	5.7	36.9	60.1	79.5	5.3	25.7	29.5	92.7	25.7	41.9	21.0	30.8	7.9	50.3	49.0	31.7	13.6	49.6	43.1	41.8
MaskCLIP	64.6	86.0	63.2	26.0	56.0	61.1	6.8	37.2	65.1	84.1	5.7	37.6	29.2	94.6	41.9	49.9	28.2	36.6	7.8	51.9	51.8	34.6	13.1	54.4	46.6	45.4
CLIP* [41]	64.1	85.3	61.1	29.1	55.6	64.1	4.8	39.1	64.7	79.4	5.8	17.8	30.6	93.3	38.6	48.9	28.8	29.0	8.1	50.4	50.5	32.8	20.5	49.5	45.9	43.9
MaskCLIP*	67.5	89.8	65.0	27.9	58.2	64.6	7.4	38.5	66.0	84.6	4.6	38.2	28.3	95.1	45.4	54.6	23.1	35.3	8.7	52.1	55.8	36.4	12.7	53.8	49.8	46.5

Table 6: Results of zero-shot image-text retrieval on Flickr30K and MS-COCO datasets. Best results in **bold**. MaskCLIP outperforms CLIP and SLIP by a large margin on both datasets.

	Training Epoch	Flickr30K						MS-COCO					
		image-to-text			text-to-image			image-to-text			text-to-image		
		R@1	R@5	R@10									
CLIP [41]	25	56.2	81.1	88.0	40.1	67.6	77.1	32.1	58.3	69.2	19.8	41.5	52.7
SLIP [39]	25	61.1	85.9	91.4	45.5	72.1	80.5	35.0	61.2	72.0	22.4	45.1	56.7
MaskCLIP	25	64.9	87.2	92.8	48.1	74.8	83.0	38.5	65.0	75.7	24.8	48.9	60.3
CLIP [41]	100	61.6	84.8	90.9	45.2	72.1	80.7	36.5	61.5	72.7	22.2	45.4	56.9
MaskCLIP	100	69.9	89.9	93.7	52.4	78.3	86.3	40.9	67.7	77.5	26.5	51.1	62.5

a strong zero-shot performance. For example, with 25 epochs training, MaskCLIP reaches 64.9% Rank@1 image-to-text accuracy on Flickr30K, outperforming CLIP with 8.7%.

5 Conclusion and Limitation

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. The resulting visual encoder shows strong transfer capability across widely adopted benchmarks for linear probing, finetuning and also zero-shot evaluation. In spite of this, one limitation of our method is that the jointly trained text encoder still learns global semantics without local semantic learning, which might hinder the capability on text encoder-dependent tasks such as zero-shot and image-text retrieval.

References

- [1] 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.
- [2] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.
- [3] 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.
- [4] 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.
- [5] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.
- [6] 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.
- [7] 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.

- [8] 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.
- [9] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.
- [10] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning, 2020.
- [11] 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.
- [12] 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.
- [13] 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.
- [14] MMsegmentation Contributors. Mmsegmentation, an open source semantic segmentation toolbox. <https://github.com/open-mmlab/msegmentation>, 2020.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] 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.
- [19] 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.
- [20] 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.
- [21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.
- [22] 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.
- [23] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.
- [24] 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.
- [25] 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.
- [26] 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.
- [27] 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.
- [28] 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.
- [29] 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.
- [30] 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.
- [31] Xiujuan 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.
- [32] 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.
- [33] 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.

- [34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [35] 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.
- [36] 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.
- [37] 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.
- [38] 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.
- [39] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. *arXiv preprint arXiv:2112.12750*, 2021.
- [40] 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.
- [41] 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.
- [42] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.
- [43] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vi-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.
- [44] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.
- [45] 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.
- [46] 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.
- [47] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luwei Zhou, and Lu Yuan. Bevt: Bert pretraining of video transformers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2022)*, 2022.
- [48] 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.
- [49] Longhui Wei, Lingxi Xie, Wengang Zhou, Houqiang Li, and Qi Tian. Mvp: Multimodality-guided visual pre-training. *arXiv preprint arXiv:2203.05175*, 2022.
- [50] 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.
- [51] 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.
- [52] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.
- [53] 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.
- [54] 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.
- [55] 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.
- [56] Chong Zhou, Chen Change Loy, and Bo Dai. Denseclip: Extract free dense labels from clip. *arXiv preprint arXiv:2112.01071*, 2021.
- [57] Luwei 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.

A Ablation study

Effect of EMA. In our experiments, we adopt a mean teacher model to provide the prediction target, instead of the student model itself. Here we explore the effect of the mean teacher (also known as EMA model). As shown in Table.8, removing mean teacher cause performance degradation in all metrics.

Table 7: Comparison between MaskCLIP and MaskCLIP without mean teacher.

Method	IN-1K			Flicker30K		ADE20K	
	0-shot	Linear	Finetune	I2T	T2I	0-shot	Finetune
MaskCLIP(Single-Stage) w/o EMA	46.5	72.0	83.5	64.2	47.9	10.3	48.8
MaskCLIP(Single-Stage)	46.6	72.5	83.8	64.9	48.1	11.2	49.6

Data Scaling. In main paper, we use at most 50M data to explore the data scaling performance of MaskCLIP, here we further extend the data to 100M with the 25 epoch training. We find the performance of MaskCLIP further improved with more training data.

Table 8: Data scaling results of MaskCLIP.

Method	Data	IN-1K			Flicker30K		ADE20K	
		0-shot	Linear	Finetune	I2T	T2I	0-shot	Finetune
MaskCLIP	LAION-20M	46.6	72.5	83.8	64.9	48.1	11.2	49.6
MaskCLIP	LAION-50M	50.9	73.4	83.9	71.4	54.5	12.6	49.9
MaskCLIP	LAION-100M	53.6	74.8	84.0	73.5	55.7	13.3	50.2

B Experiment detail

Pre-training We train our proposed MaskCLIP from scratch and training for 25 epochs, the batch size is fixed to 4096 for all the experiments. We use 32 V100 for training with 128 samples per GPU. We use the AdamW [34] optimizer with weight decay 0.5. The learning rate is set to $5e^{-4}$ with one epoch warm-up and decay to $1e^{-5}$ followed by cosine schedule. The masks used in the mask self-distillation branch are random mask with a mask ratio of 75%. The EMA weight is set to 0.999 and linearly increases to 0.9999 during the training. We pre-train all the models with a randomly sampled 20M subset from the recent image-text dataset LAION-400M [42]. We denote this subset as LAION-20M.

Zero-shot ImageNet-1K classification. For zero-shot on ImageNet-1K, we follow the prompt setting in [39] to convert the labels to text features, which contains 7 prompt templates and we use the average feature as the final label feature. We calculate the similarity between image feature and all the label features to get its zero-shot classification result.

Linear-probing ImageNet-1K classification. For linear probing, we fix the backbone and train a new linear classifier for 90 epochs. Following the setting in MAE [20], we add a batch-norm layer without learnable affine parameters before the classifier to avoid adjusting the learning rate for each model. We set the batch size to 16384 and use the LARS [52] optimizer with weight decay 0 and momentum 0.9. The learning rate is set to 6.4 and decays to 0 following the cosine schedule.

Fine-tuning ImageNet-1K classification. When fine-tuning on the ImageNet-1K dataset, we average pool the output of the last transformer of the encoder and feed it to a softmax-normalized classifier. We fine-tune 100 epochs for all the experiments, the learning rate is warmed up to 0.0006 for 20 epochs and decay to $1e^{-6}$ following the cosine schedule. Similar to recent works, we also apply the layer decayed learning rate used in [2] and we set the decay factor as 0.7. Note that we use the pure ViT architecture, *without* the techniques used in [2], such as layer scale and relative position embedding. The evaluation metric is top-1 validation accuracy of a single 224×224 crop.

Zero-shot Semantic segmentation. Here we follow the setting in DenseCLIP [56] based on the implementation from mmsegmentation [14]. For ADE20K and MS-COCO, we report the single-scale

test result with 512×512 input. For Pascal Context, we use 480×480 input. To avoid the influence of position embedding caused by changing input size, we use sliding inference with 224×224 input and stride 112. To convert the labels to text embedding, we use 85 prompt templates and use the average feature as the final label feature.

ADE20K Semantic segmentation. Here we use: UperNet [50] based on the implementation from mmsegmentaion [14]. For UperNet, we follow the settings in [2] and use AdamW [34] optimizer with initial learning rate $2e^{-4}$, weight decay of 0.05 and batch size of 16 (8 GPUs with 2 images per GPU) for 160K iterations. The learning rate warmps with 1500 iterations at the beginning and decays with a linear decay strategy. We use the layer decay [2] for the backbone and we set it as 0.6. As the ViT architecture outputs features with the same size, here we add four different scale FPNs to scale the feature map into different size. Specifically, we upsample the output feature of the 4th block $4\times$, upsample the output feature of the 6th block $2\times$, keep the output feature of the 8th block unchanged and downsample the output feature of the 12th block $2\times$. We use the default augmentation setting in mmsegmentation including random horizontal flipping, random re-scaling (ratio range [0.5, 2.0]) and random photo-metric distortion. All the models are trained with input size 512×512 . The stochastic depth is set to 0.1. When it comes to testing, we report single-scale test result.

COCO Object Detection and Instance Segmentation. We use the classical object detection framework Mask R-CNN [22] based on the implementation from mmdetection [5]. We train it the $1\times$ schedule with single-scale input (image is resized so that the shorter side is 800 pixels, while the longer side does not exceed 1333 pixels) for 12 epochs. We use AdamW [34] optimizer with a learning rate of $1e^{-4}$, weight decay of 0.05 and batch size of 16. We also use the layer decay [2] for the backbone and we set it as 0.75. The learning rate declines at the 8th and 11th epoch with decay rate being 0.1. The stochastic depth is set to 0.1. Similar to the implementation of semantic segmentation above, we also use four different scale FPNs to scale the feature map into different size.

C More visualization results.

Here we provide more visualization results on the MS-COCO val set. In most cases, our MaskCLIP gets a better feature alignment performance between image and text.

D Societal impacts

MaskCLIP is an improvement of CLIP, so it has the same societal impacts of CLIP, including some malicious usages and positive applications. Meanwhile, CLIP and MaskCLIP may suffer from some unwanted data bias, as the data used for training are roughly collected from the Internet.

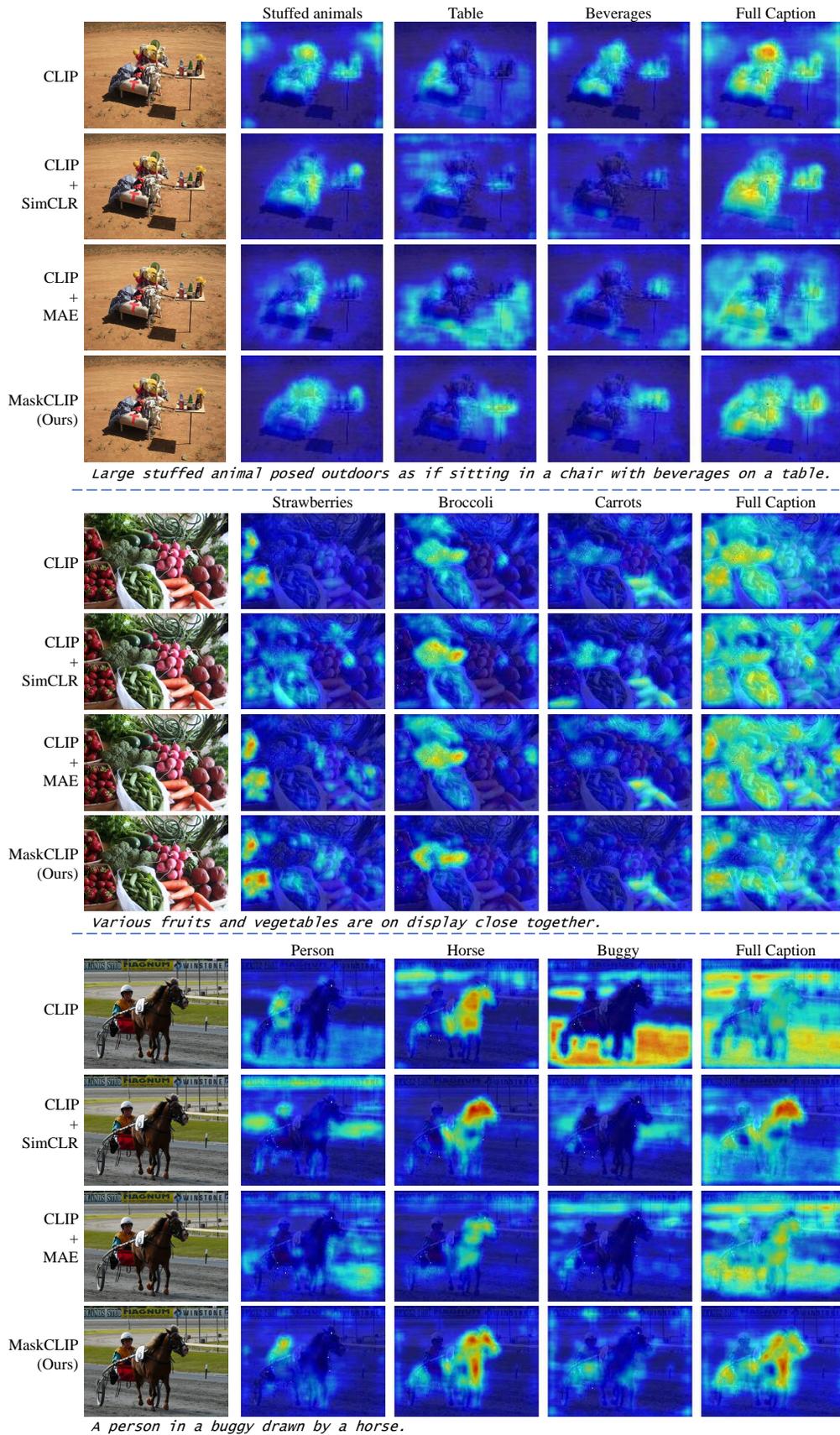


Figure 4: Visualization of the similarity between text and image features. The images and captions are from the MS-COCO val set.

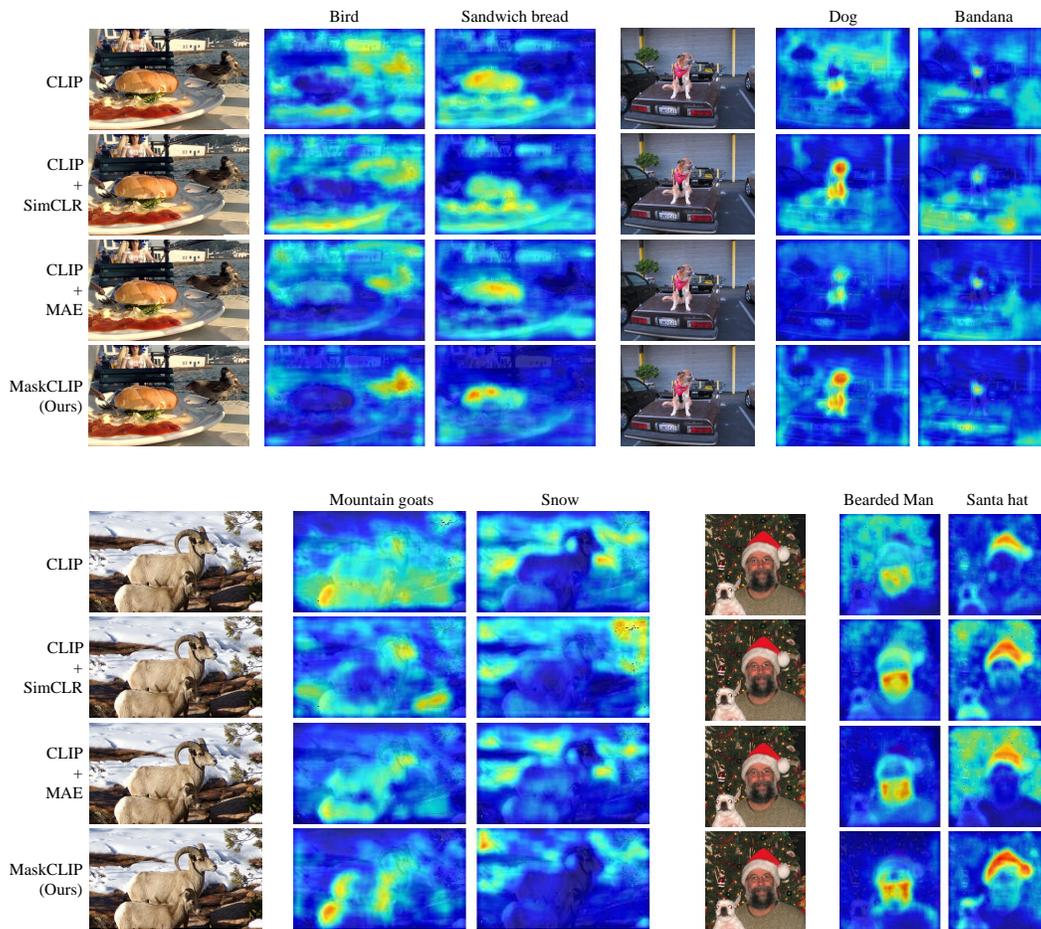


Figure 5: Visualization of the similarity between words and image features. The images and captions are from the MS-COCO val set.

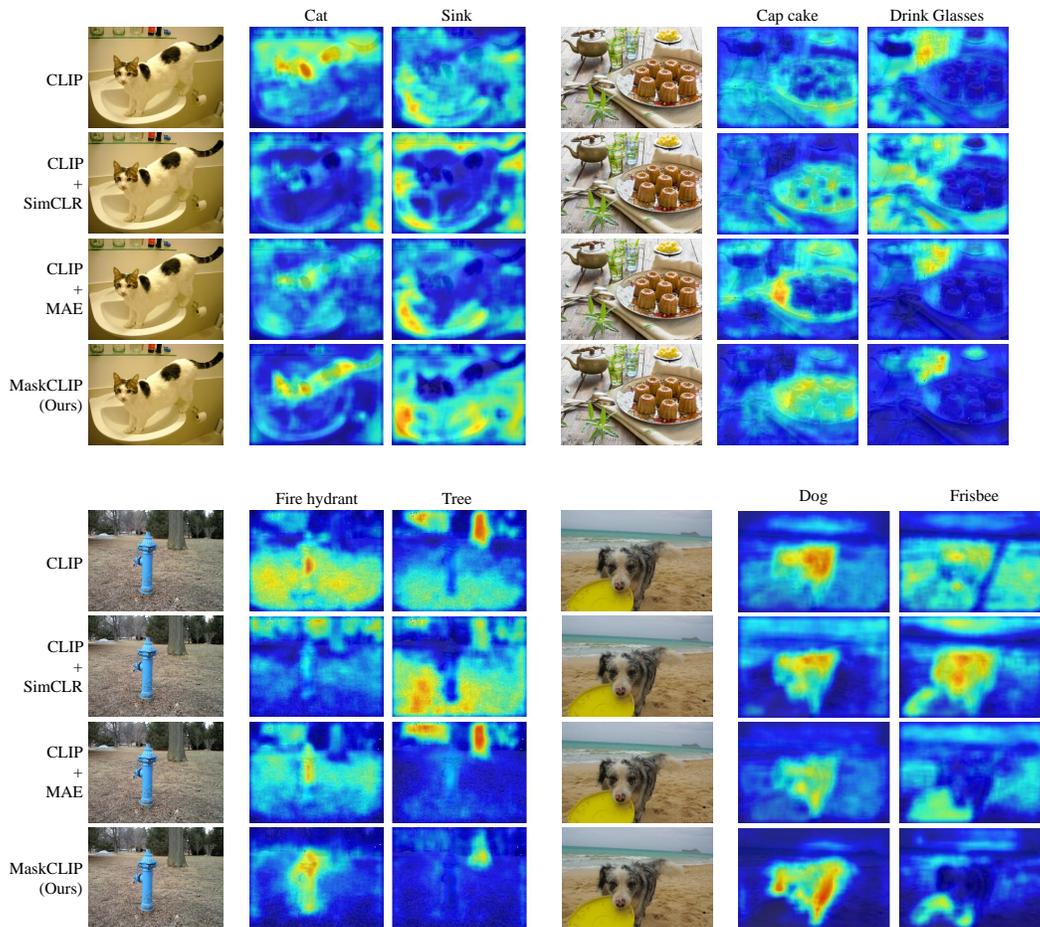


Figure 6: Visualization of the similarity between words and image features. The images and captions are from the MS-COCO val set.