Contrastive Feature Masking Open-Vocabulary Vision Transformer

Dahun Kim Anelia Angelova Weicheng Kuo

Google DeepMind

Abstract

We present Contrastive Feature Masking Vision Transformer (CFM-ViT) - an image-text pretraining methodology that achieves simultaneous learning of image- and region-level representation for open-vocabulary object detection (OVD). Our approach combines the masked autoencoder (MAE) objective into the contrastive learning objective to improve the representation for localization tasks. Unlike standard MAE, we perform reconstruction in the joint image-text embedding space, rather than the pixel space as is customary with the classical MAE method, which causes the model to better learn region-level semantics. Moreover, we introduce Positional Embedding Dropout (PED) to address scale variation between image-text pretraining and detection finetuning by randomly dropping out the positional embeddings during pretraining. PED improves detection performance and enables the use of a frozen ViT backbone as a region classifier, preventing the forgetting of open-vocabulary knowledge during detection finetuning. On LVIS open-vocabulary detection benchmark, CFM-ViT achieves a state-of-the-art 33.9 APr, surpassing the best approach by 7.6 points and achieves better zero-shot detection transfer. Finally, CFM-ViT acquires strong image-level representation, outperforming the state of the art on 8 out of 12 metrics on zero-shot image-text retrieval benchmarks.

1. Introduction

The ability to detect a vast array of objects in the real world is fundamental to computer vision and machine learning. This powers a wide range of applications from autonomous agents to search engines. Unfortunately, to date most modern object detectors rely on manually annotated regions and class labels, which is labor-intensive and impractical to scale beyond the order of 10^3 categories.

A new task called open-vocabulary detection (OVD) has been introduced to address the vocabulary limitation in object detection by using image-text pairs for training and text queries from users at test time [62]. Open-vocabulary detectors represent categories as text embeddings rather than discrete class labels, allowing them to predict objects

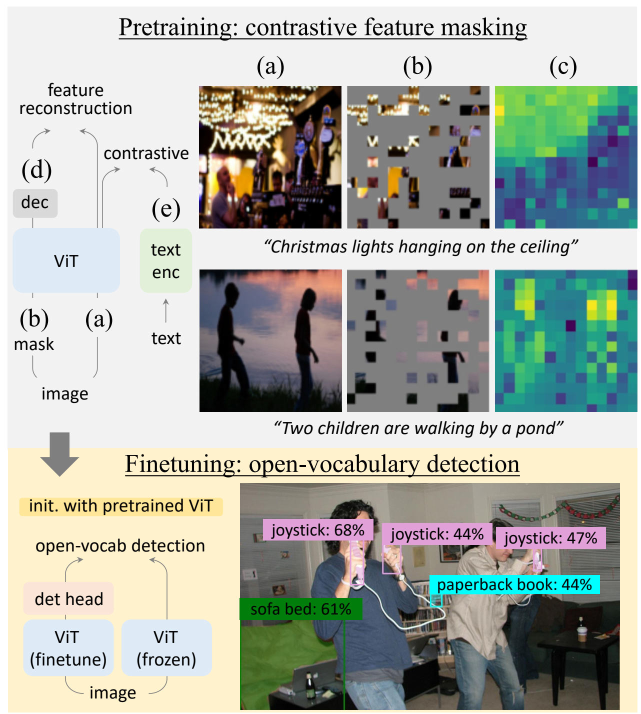

Figure 1: We propose CFM-ViT to pretrain vision transformers to capture more pixel and region information for open-vocabulary detection. CFM-ViT predicts masked contrastive features on top of the contrastive image-text pretraining. (Top) We visualize (c) the similarity map between (d) the *reconstructed* image features (see top left) and (e) the query text embedding. CFM-ViT correctly predicts the (c) whole-image semantics from (b) heavily truncated images. (Bottom) Our open-vocabulary detector exploits the frozen ViT backbone to retain pretrained knowledge and is able to detect base and novel object classes (only novel classes are shown).

unavailable during training. Various techniques, such as knowledge distillation [18, 13], weak supervision [71], self-training [68, 46, 65], and frozen backbone [32], have been suggested. Typically, CNN backbones are utilized in these approaches. As vision transformers have gained significant traction in image understanding [12, 63, 21, 3], it is crucial to explore open-vocabulary detectors based on vision transformers [39]. Moreover, to our knowledge, most current OVD research assumes the availability of pretrained Vision-

Language Models (VLMs) (*e.g*. CLIP [44]), and proposes adaptation or finetuning techniques to overcome the disparity between image-level pretraining and object-level finetuning [18, 13, 68, 65, 46]. However, as these VLMs are typically optimized for image-level tasks such as classification and retrieval, they do not adequately utilize the pixel- and region-level information during pretraining, which is crucial for downstream open-vocabulary detection.

We present CFM-ViT (Contrastive Feature Masking Vision Transformer), a simple framework to pretrain vision transformers to capture more detailed pixel/region information for open-vocabulary object detection (Fig. 1). Inspired by MAE [21], we adopt the concept of masked auto-encoding to enhance object representation during pretraining. However unlike MAE, we perform prediction in the joint image-text embedding space rather than the pixel space as an auxiliary objective to the contrastive image-text learning. This additional objective provides orthogonal signal from the contrastive learning, and benefits downstream detection task without compromising the image-level tasks. In addition, we propose Positional Embedding Dropout (PED) to address overfitting to the typically lower-resolution and object-centric pretraining data. By randomly dropping out positional embeddings during pretraining, PED aids the model to learn more robust representations that better generalize to high-res detection data. Moreover, PED enables the use of a frozen ViT encoder as an open-vocabulary region-classifier, which prevents the forgetting of open-vocabulary knowledge at detection.

We evaluate CFM-ViT on the widely used LVIS and COCO open-vocabulary detection benchmarks. Our top-performing model obtains 33.9 APr on LVIS, surpassing the previous best approach by 7.6 APr at system level. On the COCO benchmark, CFM-ViT represents the first ViT-based model and achieves a very competitive novel AP without using pseudo labels or weak supervision. Although not optimized for retrieval, CFM-ViT outperforms the state-of-the-art methods of similar or larger capacity on 8 out of 12 image-text retrieval benchmark metrics. In summary:

- We present an image-text pretraining methodology (CFM-ViT) to learn localization cues for open-vocabulary detection by contrastive feature masking.
- We propose Positional Embedding Dropout (PED) to bridge the gap between image-text pretraining and detection finetuning, which enables the use of a frozen ViT encoder to prevent the forgetting of open-vocabulary knowledge during detection finetuning.
- CFM-ViT achieves state-of-the-art APr on LVIS open-vocabulary detection benchmark, shows very competitive performance on COCO and zero-shot transfer to Objects365, and outperforms the SOTA on 8 out of 12 metrics of zero-shot image-text retrieval benchmarks.

We hope these discoveries would encourage the community to explore open-vocabulary detection from the perspective of image-text pretraining.

2. Related Works

Language-supervised open-vocabulary recognition. Learning representation for open-vocabulary recognition is a hallmark of general intelligence. Early pioneering works such as DeViSE [16] and ConSE [40] used deep convolutional networks to construct a shared image-text embedding space for zero-shot recognition. To leverage the co-occurrence of image and text in raw internet data, researchers have explored various data sources such as image tags [4, 9, 30], captions [8, 24, 47, 52], alt-texts [29, 48], image search queries [44], page title [5], or a combination of these sources [5]. From a modeling perspective, contrastive learning has become a popular paradigm because of its simplicity, scalability, and versatility in zero-shot, few-shot, and full finetuning transfer settings [43, 44, 36, 10, 34]. While most of these works focus on image-level understanding, we explore the learning of region-level information in the image-text pretraining, which is essential for open-vocabulary detection task.

Self-supervised object representation learning. Scaling up annotation for detection presents a significant challenge. As a result, many efforts have been made to learn object representations in a self-supervised manner. These approaches can be broadly categorized as contrastive or generative. These contrastive approaches typically use sliding windows [56], object proposals [54, 25], or point samples [1] for pixel or region-level contrastive learning. Generative methods use masked image modeling with reconstruction targets such as pixels [21], low-level [3, 53] / high-level image features [6, 70], or combine with the contrastive objective [27]. By learning to restore masked images, the model needs to learn about objects and regions. However, although these self-supervised methods are suited for localization tasks, they lack the necessary image-text learning for open-vocabulary recognition. Some recent works [55, 42, 64, 26, 14] utilize off-the-shelf CLIP features [44] as prediction targets to enhance masked image modeling by two-stage training. In this work, we propose a novel approach to combine generative self-supervised learning jointly with contrastive image-text learning in a single end-to-end training stage. While some concurrent works have explored similar objectives for zero-shot image-level tasks or fully supervised finetuning [11, 57, 51], our focus is on open-vocabulary detection.

Open-vocabulary object detection and segmentation. Zero-shot detection aims to enhance detection models beyond their limited training categories by aligning region visual representation and category word embeddings [2, 45, 7, 66] or generating visual features with a generative model [20, 72]. Open-vocabulary detection [62] improves upon zero-shot detection by incorporating image-text supervision about the novel categories. With the advent of image-text pretraining, numerous studies have explored adapting these pretrained models to open-vocabulary detection and segmentation [18, 68, 17, 33, 69]. For instance, ViLD [18] distills image-text knowledge into the detector, while DetPro [13] improves ViLD by category prompt optimization. Additionally, region-text self-training has been demonstrated on image caption data [68], classification data [46], and unlabeled data [65]. Phrase grounding [35], weak supervision [71], and frozen model [32] approaches have also been explored. Most methods rely on CNN backbones, but vision transformers are gaining momentum [39, 69]. While previous studies have focused on finetuning or adaptation strategies for pretrained models, ours seeks to improve the image-text pretraining by predicting the masked representation of vision transformer.

3. Method

We tackle the problem of open-vocabulary object detection. During training, the model can access the detection labels of base categories, but at the inference phase, it must be able to detect objects from a set of novel categories. To achieve this, we utilize pretrained vision and language models (VLMs) following previous works [18, 68, 32]. However, instead of taking off-the-shelf pretrained VLM, we demonstrate how to better pretrain VLMs with vision transformers [12] for open-vocabulary detection.

3.1. Preliminaries: Overall Pipeline

Pretraining. We adopt a dual-encoder image-text contrastive model widely used in existing works [44, 29]. The image embeddings $\{v\}$ and text embeddings $\{l\}$ are obtained by global average pooling at the last layers of image and text encoders. The cosine similarity of the embeddings in batch B, scaled by a learnable temperature τ are the input to the InfoNCE loss [41, 44]. The image-to-text (I2T) contrastive loss is formulated as:

$$L_{\text{I2T}} = -\frac{1}{B}\sum_{i=1}^{B}\log\left(\frac{\exp(v_i l_i/\tau)}{\sum_{j=1}^{B}\exp(v_i l_j/\tau)}\right). \tag{1}$$

The text-to-image (T2I) contrastive loss is symmetrical with the I2T loss by exchanging the inner/outer summation loops. The total contrastive loss L_{con} is obtained by $L_{con} = (L_{\text{I2T}} + L_{\text{T2I}})/2$.

Downstream open-vocabulary detection. Our open-vocabulary detection algorithm follows existing works [62, 18, 32]. At training, for each detected region i, its region embedding is the RoI-Align feature. The detection score p_i is the cosine similarity between the region embedding and text embeddings of C_B followed by a softmax. Note the text embeddings are computed from the same text encoder from the image-text pretraining. At test time, the text embeddings are expanded from the C_B to $C_B \cup C_N$ plus the "background" embedding. We also extract VLM embedding of region i by RoI-Align at the last feature map of the ViT backbone. The VLM score z_i is the cosine similarity with the $C_B \cup C_N$ text embeddings. Similarly, the detection score p_i is now computed with $C_B \cup C_N$ text embeddings.

An object detector for open-vocabulary scenarios is trained on the labels of base categories C_B, but must be capable of detecting the union of base and novel categories ($C_B \cup C_N$) at test time. Following existing works [62, 18], we replace the fixed-size classifier layer with the text embeddings of base categories. The same text encoder from the image-text pretraining is used to compute the text embeddings to maintain the pretrained open-vocabulary knowledge. The "background" phrase represents the background category, and the proposals not matched to any C_B annotations are labeled as background.

The ensemble open-vocabulary detection score s_i^{ens} is obtained by geometric means [18, 32]:

$$s_i^{\text{ens}} = \begin{cases} z_i^{(1-\alpha)} \cdot p_i^{\alpha} & \text{if } i \in C_B \\ z_i^{(1-\beta)} \cdot p_i^{\beta} & \text{if } i \in C_N \end{cases} \tag{2}$$

, where $\alpha, \beta \in [0, 1]$ control the weights for base and novel categories. The background score comes directly from the detection score p_i, because the VLM score with "background" phrase tends to be not as reliable.

3.2. Contrastive Feature Masking

Our method performs reconstruction in the joint image-text embedding space (see Fig. 2-left) as an auxiliary objective to the contrastive image-text learning (in Sec. 3.1).

Masked feature reconstruction. Following MAE [22], we randomly mask a large portion of image tokens (*e.g.*, mask ratio 75%) for representation learning. However unlike MAE, we predict the joint image-text embedding instead of the raw pixels to encourage better learning of semantics. Specifically, the output features $\{f\}$ of the contrastive image encoder before the global average pooling is our reconstruction target. We use the cosine distance between the reconstructed features $\{\hat{f}\}$ and unmasked image features $\{f\}$ as loss function. Let M be the set of masked patch indices, and our reconstruction loss L_{rec} is computed only on the masked tokens as:

$$L_{rec} = 1 - \frac{1}{B}\sum_{i=1}^{B}\left(\frac{1}{|M|}\sum_{k\in M}\frac{f \cdot \texttt{sg}(\hat{f})}{\|f\| \cdot \|\texttt{sg}(\hat{f})\|}\right), \tag{3}$$

where $|M|$ is the number of masked tokens and $\texttt{sg}$ denotes stop gradient. The total CFM-ViT loss is $L_{con} + L_{rec}$.

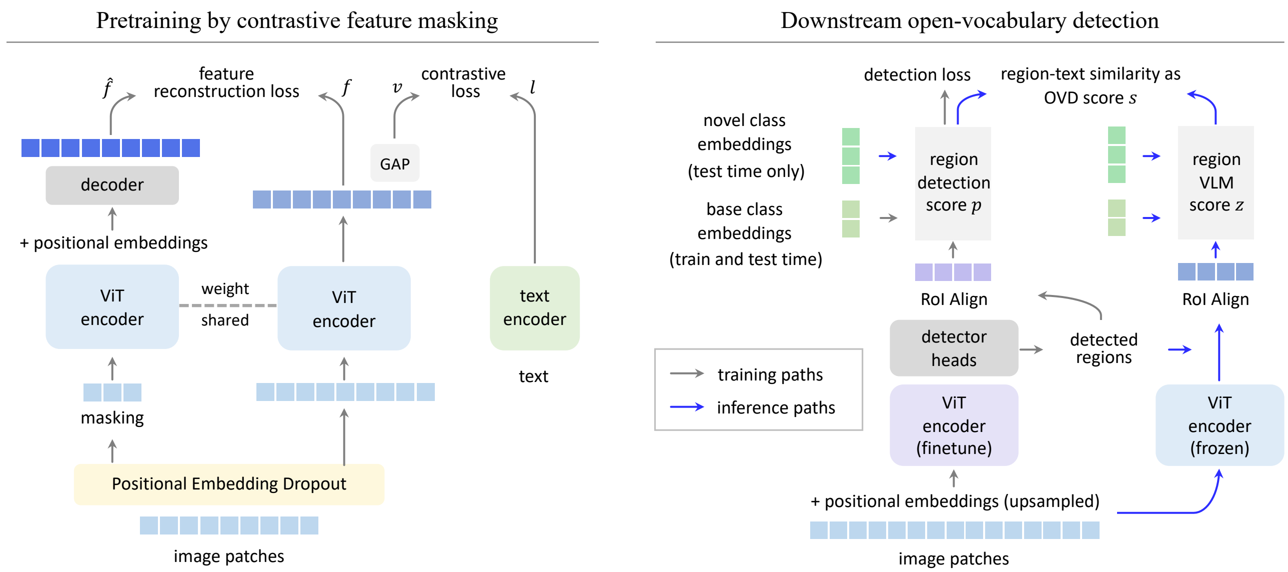

Figure 2: **CFM-ViT architecture:** We present both the image-text pretraining (left) and open-vocabulary detection finetuning (right) architecture of CFM-ViT. (Left) Building upon contrastive learning, we learn to reconstruct the masked tokens in the joint image-text embedding space. In addition, we propose Positional Embedding Dropout (PED) which randomly masks out the whole PE during pretraining to mitigate overfitting to the low-res positional embeddings, thus adapting better to the high-res downstream detection task. (Right) The open-vocabulary detector is initialized with the pretrained ViT backbone during finetuning. The detected region embeddings match with the cached category embeddings to compute the region scores. At inference, we exploit the frozen ViT backbone to obtain the VLM score z, which is combined with the detection score p into the open-vocabulary detection score s (Best viewed in color).

Our reconstruction encoder is identical (weight-shared) to the contrastive image encoder, but applied only on the visible, unmasked tokens (*e.g.*, 25%). The decoder takes the encoded visible tokens and learnable `[mask]` tokens added with positional embeddings.

Faster training by contrastive branch masking. The feature reconstruction branch adds a computation burden (*e.g.* 25%) to the pretraining depending on the masking ratio (*e.g.* 75%). We note that this cost can be waived by feeding only the masked tokens (M) to the contrastive branch, so that the input patches to the contrastive and reconstruction encoders are mutually exclusive, and yields the same reconstruction target $\{\hat{f}_{k \in M}\}$. Our ablation study in Table 5c shows that this technique maintains the training efficiency of contrastive learning, while still achieves significant gains over the baseline in open-vocabulary detection.

Positional embedding dropout. In vision transformer encoder, positional embeddings are added to all tokens after the first patchifying layer to provide the location of each patch in the image. While the positional embeddings work well for image classification/retrieval, it tends to overfit to the lower-resolution object-centric images, and struggle with higher-resolution images typically used by detection task. In addition, the recognition of objects in detection occurs at region- rather than image-level (*e.g.* see VLM scores z_i for region i in Sec. 3.1), which causes difficulty for the positional embeddings trained only for image-level task.

We propose a simple yet effective technique called Positional Embedding Dropout (PED) to address this problem by randomly masking out the whole positional embeddings during training (*e.g.*, with a probability 0.5). This teaches the model not to rely heavily on the positional embeddings and thus can process the high-res images and perform better region classification. PED not only outperforms both the baseline and 'no positional embeddings' variants, but enables the use of *frozen* vision transformer to achieve further improvement in open-vocabulary detection.

3.3. Open-vocabulary Detection

An object detector for open-vocabulary scenarios is trained on the labels of base categories C_B, but must be capable of detecting the union of base and novel categories ($C_B \cup C_N$) at test time (see Sec. 3.1 and Fig. 2-right).

Baseline architecture. Our detector adopts the simple feature pyramid and windowed attention to handle higher resolution images as proposed in ViTDet [37], and employs Mask R-CNN heads and class-agnostic box regression and mask heads as in [13, 18, 62, 68, 32]. In addition, we leverage a recent novel object proposal method [31] by replacing the binary classification in the RPN with the centerness-based objectness. The predicted objectness score o_i is combined into the final OVD score as $s_i^{\text{OVD}} = o_i \cdot s_i^{\text{ens}}$.

Our detector backbone is initialized with the pretrained ViT in the VLM from Sec. 3.2, and is finetuned together

method	pretrained model	detector backbone	$\mathbf{AP}_r$	AP
ConvNet based:				
DetPro-Cascade [13]	ViT-B/32	R-50	20.0	27.0
Detic-CN2 [71]	ViT-B/32	R-50	24.6	32.4
RegionCLIP [68]	R-50x4	R-50x4	22.0	32.3
ViLD-Ens [18]	ViT-B/32	R-152	18.7	26.0
ViLD-Ens [18]	ViT-L/14	EffNet-B7	21.7	29.6
ViLD-Ens [18]	EffNet-B7	EffNet-B7	26.3	29.3
VL-PLM [65]	ViT-B/32	R-50	17.2	27.0
OV-DETR [61]	ViT-B/32	R-50	17.4	26.6
Rasheed *et al.* [46]	ViT-B/32	R-50	21.1	25.9
PromptDet [15]	ViT-B/32	R-50	21.4	25.3
ViT based:				
OWL-ViT [39]	ViT-H/14	ViT-H/14	23.3*	35.3*
OWL-ViT [39]	ViT-L/14	ViT-L/14	25.6*	34.7*
CFM-ViT (ours)	ViT-B/16	ViT-B/16	**29.6***	33.8*
CFM-ViT (ours)	ViT-L/16	ViT-L/16	**35.6***	38.5*
CFM-ViT (ours)	ViT-B/16	ViT-B/16	**28.8**	32.0
CFM-ViT (ours)	ViT-L/16	ViT-L/16	**33.9**	36.6

Table 1: **LVIS open-vocabulary object detection.** CFM-ViT outperforms the best existing approach by **+7.6** $\mathbf{AP}_r$, and the other ViT-based approach [39] by **+10.0** $\mathbf{AP}_r$ using the same backbone. $*$: reports box AP.

with the newly added detector heads. Note we do *not* apply positional embedding dropout (PED) during finetuning as the location information is critical in detection.

Backbone learning rate. As the pretrained knowledge in the backbone is critical in recognizing novel categories, it is important to set the backbone learning rate so as to prevent forgetting in the finetuning phase. On the other hand, entirely freezing the backbone limits the ability to adapt to detection tasks. We find that setting the backbone learning rate lower (*e.g.*, 0.5×) than the rest of the detector layers shows advantage in the trade-off. After the detection training is done, we explore using the frozen ViT backbone at test time, as described next.

Frozen backbone inference While the ViT backbone adapts to the detection tasks, it tends to forget some of the pretrained open-vocabulary knowledge. Therefore, for inference, we propose to use a separate frozen ViT backbone as an open-vocabulary region classifier. Specifically, we use the frozen backbone instead of the finetuned backbone when computing the region VLM score z_i (Sec. 3.1). We find it important for the frozen ViT to be pretrained with our positional embedding dropout (PED), to serve as a strong zero-shot region classifier. We show by experiments that incorporating the PED pretraining and frozen backbone inference provides large gains in open-vocabulary detection.

4. Experimental Results

Pretraining setup. For the image-text pretraining, we use the widely-used ViT-B/16 and ViT-L/16 as the image encoder, with an input image size of 224. We use the fixed 2D sinusoidal positional embeddings, and apply Positional Embedding Dropout (PED) with a drop probability of 0.5. The image embedding is obtained by global average pooling at the last ViT layer. The text encoder is a 12-layer Transformer as in [44, 59], with the input sequences truncated to a fixed length of 64 tokens. The L2-normalized image and text embeddings and a learnable scaling temperature are the input to the InfoNCE contrastive loss [44].

method	pretrained model	detector backbone	**novel AP**	AP
ConvNet based:				
ViLD [18]	ViT-B/32	R-50	27.6	51.3
OV-DETR [61]	ViT-B/32	R-50	29.4	52.7
w/ pseudo box labels:				
XPM *et al.* [28]	R-50	R-50	27.0	41.2
RegionCLIP [68] †	R-50x4	R-50x4	39.3	55.7
PromptDet [15]	ViT-B/32	R-50	26.6	50.6
VL-PLM [65]	ViT-B/32	R-50	34.4	53.5
Rasheed *et al.* [46] ‡	ViT-B/32	R-50	36.9	51.5
w/ weak supervision:				
Detic-CN2 [71]	ViT-B/32	R-50	24.6	32.4
ViT based:*				
CFM-ViT (ours)	ViT-B/16	ViT-B/16	30.8	42.4
CFM-ViT (ours)	ViT-L/16	ViT-L/16	34.1	46.0

Table 2: **COCO open-vocabulary object detection (box AP50).** CFM-ViT represents the first ViT-based approach and demonstrates a very competitive novel AP without using pseudo labeling or weak supervision. †: RegionCLIP uses an off-the-shelf RPN during its pretraining. ‡: Rasheed *et al.* uses an external MViT detector [38] during pretraining. *: The other ViT-based method [39] report their results on LVIS only.

Our feature reconstruction decoder is a 2-layer ViT, unlike the 8-layer counterpart of MAE [22] designed for raw pixel reconstruction. The reconstruction loss is cosine distance, scaled by a loss coefficient 2.0, and is added to the contrastive loss. We use the same image-text dataset as Jia et al. [29]. Unless noted, we use a batch size of 4k for ablation and 16k for comparisons, and train for 500k iterations using the AdamW optimizer with an initial learning rate (LR) of 5e-4 and linear LR decay. We use 10k warm-up iterations and a weight decay of 0.01.

Detection finetuning setup. We train our model on base categories C_B with an image size of 1024×1024. The positional embeddings (PE) are bilinearly interpolated to fit the higher resolution. We do *not* apply PE Dropout during the detection training, and set a lower learning rate for the backbone (*e.g.*, 0.5 ×) compared to the rest of the model. We utilize CLIP templates [44] and take the average text embeddings of each category. We use a batch size 128, the SGD optimizer with momentum 0.9, an initial learning rate of 0.18/0.02 and train for 36.8k/11.3k iterations on LVIS/COCO datasets.

method	image encoder size	Flickr30K (1K test set)						MS COCO (5K test set)					
		image-to-text			text-to-image			image-to-text			text-to-image		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R10	R@1	R@5	R@10
CLIP [44]	302M	88.0	98.7	99.4	68.7	90.6	95.2	58.4	81.5	88.1	37.8	62.4	72.2
ALIGN [29]	480M	88.6	98.7	99.7	75.7	93.8	96.8	58.6	83.0	89.7	45.6	69.8	78.6
FLAVA [50]	86M	67.7	94.0	-	65.2	89.4	-	42.7	76.8	-	38.4	67.5	-
FILIP [58]	302M	89.8	99.2	99.8	75.0	93.4	96.3	61.3	84.3	90.4	45.9	70.6	79.3
Florence [60]	637M	90.9	99.1	-	76.7	93.6	-	64.7	85.9	-	47.2	71.4	-
CoCa-Large [59]	303M	91.4	**99.2**	**99.9**	79.0	95.1	97.4	65.4	85.6	91.4	**50.1**	**73.8**	**81.8**
CFM-ViT (ours)	303M	**91.7**	99.0	**99.9**	**79.6**	**95.6**	**97.7**	**66.4**	**86.1**	**91.5**	49.8	73.5	81.6

Table 3: **Zero-shot image-text retrieval results on Flickr30K and COCO benchmarks.** We evaluate our pretrained model compared to other methods. We outperform the state-of-the-art CoCa-Large with the same backbone in 8 out of 12 metrics.

4.1. Main Results

LVIS benchmark. We compare with other methods on the LVIS [19] open-vocabulary detection benchmark which contains a diverse set of 1203 object categories. The base categories C_B for training are the 'frequent' and 'common' categories, and novel categories C_N are the 'rare' categories which are held out for testing, as in [18, 67, 13]. The main metric is mask AP$_r$, and we report the mean over three runs following [18] for reproducibility.

Table 1 reports that the best CFM-ViT model achieves 33.9 AP$_r$, a significant improvement over the best existing ViT-based method OWL-ViT [39] by **+10.0** AP$_r$. Remarkably, CFM-ViT using a smaller ViT-B/16 backbone outperforms OWL-ViT with ViT-L/14 by **+4.0** AP$_r$. Furthermore, compared to the current best approach ViLD-Ens with EffNet-B7 backbone, CFM-ViT achieves a **+7.6** AP$_r$ improvement. Notably, CFM-ViT has a simple finetuning recipe using only vanilla detection losses [23], without the use of long-tail recognition losses [39, 68, 71], knowledge distillation [18, 13], weak supervision [71], or pseudo box/mask labels [68, 65, 46], all of which are common among current open-vocabulary detection methods.

COCO benchmark. We present the comparison on the COCO open-vocabulary detection benchmark. This setup uses 48 base categories for training and 17 novel categories for testing [18]. The main metric is AP50 of novel categories ('novel AP'). Due to fewer training categories, the CFM-ViT model has a tendency to overfit to these categories using only the vanilla detection losses. This is because CFM-ViT do not use any auxiliary objectives such as pseudo box/mask labels [28, 15, 68, 65, 46], knowledge distillation [18, 13], weak supervision [71] to counter-balance overfitting on this benchmark. However, Table 2 shows that CFM-ViT is still very competitive among existing methods leveraging auxiliary objectives. Moreover, CFM-ViT represents the first ViT-based method on this benchmark, as the other ViT-based [39] approach only benchmarks on LVIS.

Zero-shot Image-Text Retrieval. In addition to our main evaluation on the region-level open-vocabulary detection, we evaluate our *image-level* representation in zero-shot image-text retrieval. We take the same CFM-ViT model as in the last row of Table 1 (ViT-L, batch size 16k) and continue the pretraining on higher resolution, *e.g.*, 448, for extra 40K iterations, following the standard protocol [29, 59].

Table 3 shows our comparison with other dual-encoder methods on Flickr30K and MS COCO benchmarks. CFM-ViT outperforms state-of-the-art methods of similar or larger model size, on 8 out of 12 metrics.

method	backbone	AP	AP$_{50}$	AP$_{75}$
supervised [18]	R-50	25.6	38.6	28.0
ViLD [18]	R-50	11.8	18.2	12.6
DetPro [13]	R-50	12.1	18.8	12.9
CFM-ViT (ours)	ViT-B/16	**15.9**	**24.6**	**17.4**
CFM-ViT (ours)	ViT-L/16	**18.7**	**28.9**	**20.3**

Table 4: **Transfer detection on Objects365 (Box APs).** All models are trained on the LVIS base categories and tested on Objects365 dataset, without finetuning.

Zero-shot Transfer Detection. To assess CFM-ViT's ability to generalize in zero-shot transfer detection, we test its performance on Objects365-v1 validation split [49]. We use the same detector trained on LVIS base categories (Table 1) and replace LVIS with Objects365 vocabulary embeddings for transfer detection without finetuning [18, 13]. We assume all categories are novel and set α, β=(0.0, 0.65) in Eq. (2). Our best model achieves 18.7 AP, outperforming ViLD by +6.9 AP and DetPro by +5.6 AP, as shown in Table 4. Given the different backbone capacity (R50 vs ViT), this comparison mainly serves to demonstrate that CFM-ViT can achieve strong cross-dataset generalization.

4.2. Ablation Study

We ablate the design of CFM-ViT's pretraining and open-vocabulary detector. We evaluate on the LVIS open-vocabulary detection benchmark. The image encoder is ViT-L/16, and contrastive batch size is 4k by default.

Masked feature reconstruction. Table 5a ablates the proposed masked image-text pretraining (Sec. 3.2). The proposed masked feature reconstruction offers a clear ben-

pretraining method	AP_r	AP
baseline	27.4	30.4
w/ feat recon.	30.7 (+3.3)	34.0
w/ pixel recon.	27.1	31.3
w/ 1st-layer feat recon.	27.2	30.8

(a) **Masked reconstruction**. 'baseline' is the contrastive image-text pretraining. Our proposed masked feature reconstruction improves by +3.3 AP_r. Reconstruction in the raw pixel space or the first-layer feature space shows no benefit.

pretraining method	AP_r	AP
baseline	27.4	30.4
w/ PED	28.5 (+1.1)	31.9
w/ feat recon. + PED	31.2 (+3.8)	33.7
w/ no PE	25.8	29.5
w/ feat recon. + no PE	27.7	31.9

(b) **Positional Embedding Dropout (PED)** improves the baseline by 1.1 AP_r. It achieves a further gain of +2.7 when used with masked feature reconstruction. PED outperforms 'no PE' by 3.5 / 1.6 AP_r with/without feature reconstruction

contr. / recon.	FLOPs	AP_r	AP
100% / 0%	1.00×	27.4	30.4
100% / 25%	1.23×	30.7	34.0
100% / 50%	1.44×	29.9	33.1
75% / 25%	**1.01**×	30.4	33.9

(c) **Masking contrastive branch** recovers the training efficiency with little or no performance drop, outperforming the baseline by +3.0 AP_r.

bblr	AP_r	AP
0.0	9.5	11.4
0.1	25.8	28.5
0.5	**27.4**	30.4
1.0	26.0	30.2

(d) **Backbone finetuning lr ratio** (bblr) w.r.t. added detector layers.

	w/ PED	AP_r	AP
baseline		27.4 → 24.6 (-2.8)	30.4 → 30.3
w/ feat-recon.		30.7 → 27.1 (-3.8)	34.0 → 33.4
baseline	✓	28.5 → 30.5 (+2.0)	31.9 → 31.8
w/ feat-recon	✓	31.2 → 32.5 (+1.3)	33.7 → 34.1

(e) **Frozen backbone inference.** When using standard positional embeddings, it underperforms the finetuned encoder. In contrast, with the encoder pretrained with PED, the frozen backbone inference surpasses the finetuned counterpart by +2.0 and +1.3 AP_r.

model	batch	AP_r	AP
B/16	4k	24.1 → 26.8 (+2.7)	27.6 → 30.2
B/16	16k	26.4 → 28.8 (+2.4)	30.3 → 33,5
L/16	4k	27.4 → 32.5 (+5.1)	30.4 → 34.1
L/16	16k	30.5 → 33.9 (+3.4)	35.9 → 36.6

(f) **Scalabiltiy:** The benefit of **'baseline → CFM-ViT'** across different model and contrastive batch sizes. It improves the baselines by +2.4 to +5.1 AP_r.

Table 5: **Ablation studies** on LVIS open-vocabulary detection benchmark. We train on base ('frequent' + 'common') categories, test on novel ('rare') categories, and report AP_r. We use ViT-L/16 backbone and contrastive batch size 4k unless otherwise noted.

efit of +3.3 AP_r over the contrastive image-text pretraining baseline. In this case, the feature reconstruction target is the output features of the image encoder. We compare with other reconstruction targets: normalized image pixels [22] and the features from the first patchifying layer. We observe that neither improve over the baseline, likely because the contrastive pretraining sets a strong baseline representation [18, 10, 32]. In contrast, the proposed masked feature reconstruction clearly improves upon the strong baseline and shows advantage in open-vocabulary detection.

Positional embedding dropout. In Table 5b, we ablate the positional embedding dropout ('PED'). PED brings a gain of +1.1 AP_r over the baseline (PE without dropout). This shows that PED effectively reduces overfitting to the low-res whole-image PE during pretraining, thus adapting better to the high-res detection task through finetuning. In addition, PED achieves further gain of +2.7 when used together with masked feature reconstruction. We compare PED with another baseline which uses no positional embeddings in the ViT encoder ('no PE'). The PED method outperforms the 'no PE' baseline by 3.5 / 1.6 AP_r with/without feature reconstruction. We note that the positional embeddings in the reconstruction decoder [22] is always kept. Finally, PED allows the use of the *frozen* backbone as a strong *region* classifier as shown in Table 5e.

Faster training by masking contrastive branch. Table 5c studies image masking ratios of the contrastive and reconstruction encoders. By default, we apply our contrastive encoder on intact images during training, *i.e.* 100% tokens. Adding the reconstruction tower with 25% input tokens results in 1.23× more training cost. To maintain the training efficiency, we explore feeding only 75% tokens to the contrastive encoder that are mutually exclusive from the reconstruction branch inputs. This masking technique fully recovers the training efficiency with little or no accuracy loss, outperforming the baseline by +3.0 AP_r.

	Flickr30K		MS COCO	
	I2T	T2I	I2T	T2I
baseline	86.0	72.3	59.3	43.4
w/ PED	86.1	72.5	59.1	43.2
w/ feat recon. + PED	87.0	73.6	60.1	44.2

Table 6: **Pretraining evaluation on zero-shot image-text retrieval (Recall@1).** We evaluate the image-level representation of our pretrained model on Flickr30k and COCO retrieval tasks. We ablate the positional embedding dropout (PED) and adding masked feature reconstruction. ViT-L/16 and batch size 4k is used.

Backbone learning rate ratio. CFM-ViT requires the retention of pretrained knowledge in the backbone to recognize novel categories. Table 5d reports the advantage to set the backbone learning rate lower than the rest of the detector during the finetuning, with a ratio 0.5× being the optimal value. Higher ratios lead to forgetting, while lower ratios limit the ability to adapt to the detection task.

Frozen backbone inference. Our ablation studies so far do *not* involve frozen backbone inference. All ablations use the *finetuned* ViT backbone to compute the VLM scores (p_i in Sec. 3.1 and Eq. (2)). In Table 5e, we assess the efficacy of the frozen backbone as a region classifier by substituting the finetuned ViT encoder with a frozen ViT en-

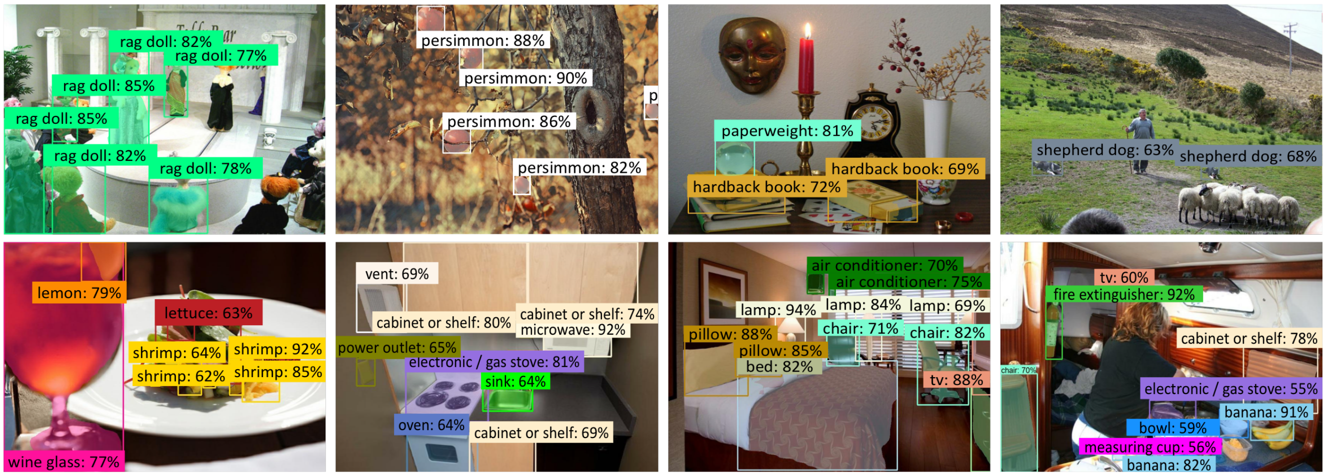

Figure 3: **Qualitative results on LVIS novel categories (top) and Objects365 zero-shot transfer detection (bottom).** For LVIS results, we only show the novel categories for clarity. CFM-ViT detects many novel categories such as *rag doll, persimmon, paperweight, hardback book, shepherd dog* on LVIS, and *shrimp, power outlet* on Objects365.

coder and analyze the performance (see the rightmost part of Fig. 2). Our experiments show that the frozen backbone underperforms the finetuned encoder when using standard positional embeddings, which applies to both the baseline with and without feature reconstruction loss. However, we find that pretraining the ViT encoder with positional embedding dropout (PED) leads to signficantly improved performance with frozen backbone, surpassing thoese of the finetuned backbone by +2.0/+1.3 APr, without/with feature reconstruction loss. This result demonstrates the efficacy of PED in reducing the domain gap between contrastive pretraining and detection finetuning, thus improving zero-shot region classification. Combined with feature reconstruction, our full method achieves an overall improvement of +5.1 APr over the baseline.

Model size and batch size. Table 5f studies the effect of model size and batch size in CFM-ViT pretraining on the downstream open-vocabulary detection. We observe that increasing the batch size from 4k to 16k leads to an improvement of +2.7 / 1.4 APr for both ViT-B/L, while upgrading from ViT-B to ViT-L results in an improvement of +5.9 / 5.6 APr for both batch sizes. These results align with observations from the contrastive learning literature [44, 29, 43] that larger batch sizes and model sizes are both highly beneficial. Importantly, we find that CFM-ViT consistently outperforms the baseline by +2.4 to +5.1 APr, across all batch and model sizes tested, further demonstrating its efficacy.

Image-text retrieval. In addition to ablations on open-vocabulary detection, we investigate the effects of positional embedding dropout and masked feature reconstruction on zero-shot *image-level* retrieval, and report the results in terms of Recall@1 metrics on Flickr30K and MS COCO datasets. Table 6 shows that positional embedding dropout effectively preserves the quality of image-level representation, while masked feature reconstruction yields an average improvement of 1% Recall@1 across all metrics.

4.3. Visualizations

Feature reconstruction. In Fig. 1, we show our feature reconstruction results from our pretraining (Sec. 3.2). For visualization, we compute the similarity map (c) between the *reconstructed* image features (d), and a query text embedding (e). We observe that the learned feature reconstructions are semantically plausible with respect to the queried image-text pairs.

Open-vocabulary detection outputs. In Fig. 3, we visualize our CFM-ViT outputs on LVIS novel categories (top row) and zero-shot transfer detection on Objects365 (bottom row). For both visualizations, we use the same model as in the last row of Table 1, which is trained on the LVIS base categories. On both datasets, CFM-ViT is able to detect many novel categories unavailable during training.

5. Conclusion

We introduce Contrastive Feature Masking Vision Transformer (CFM-ViT) which imbues the image-text pretraining with pixel/region-level semantics for open-vocabulary object detection. By using feature construction and positional embedding dropout, CFM-ViT is simple and scalable, outperforming the state-of-the-art on LVIS open-vocabulary detection benchmark by large margins, and shows very competitive performance on COCO benchmark and zero-shot transfer to Objects365. In addition, CFM-ViT outperforms the state-of-the-art on 8 out of 12 metrics of zero-shot image-text retrieval benchmarks. We hope CFM-ViT would inspire the community to explore image-text pretraining for open-vocabulary detection.

References

[1] Yutong Bai, Xinlei Chen, Alexander Kirillov, Alan Yuille, and Alexander C. Berg. Point-level region contrast for object detection pre-training. In *CVPR*, pages 16061–16070, June 2022. 2

[2] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *ECCV*, 2018. 3

[3] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 1, 2

[4] Xinlei Chen and Abhinav Gupta. Webly supervised learning of convolutional networks. In *ICCV*, 2015. 2

[5] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 2

[6] Yabo Chen, Yuchen Liu, Dongsheng Jiang, Xiaopeng Zhang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Sdae: Self-distillated masked autoencoder. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*, pages 108–124. Springer, 2022. 2

[7] Berkan Demirel, Ramazan Gokberk Cinbis, and Nazli Ikizler-Cinbis. Zero-shot object detection by hybrid region embedding. In *BMVC*, 2018. 3

[8] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *CVPR*, 2021. 2

[9] Santosh K Divvala, Ali Farhadi, and Carlos Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *CVPR*, 2014. 2

[10] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Shuyang Gu, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Clip itself is a strong fine-tuner: Achieving 85.7% and 88.0% top-1 accuracy with vit-b and vit-l on imagenet. *arXiv preprint arXiv:2212.06138*, 2022. 2, 7

[11] Xiaoyi Dong, Yinglin Zheng, Jianmin Bao, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. *arXiv preprint arXiv:2208.12262*, 2022. 2

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 3

[13] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *CVPR*, 2022. 1, 2, 3, 4, 5, 6

[14] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*, 2022. 2

[15] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Promptdet: Towards open-vocabulary detection using uncurated images. In *European Conference on Computer Vision*, pages 701–717. Springer, 2022. 5, 6

[16] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NeurIPS*, 2013. 2

[17] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022. 3

[18] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022. 1, 2, 3, 4, 5, 6, 7

[19] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 6

[20] Nasir Hayat, Munawar Hayat, Shafin Rahman, Salman Khan, Syed Waqas Zamir, and Fahad Shahbaz Khan. Synthesizing the unseen for zero-shot object detection. In *ACCV*, 2020. 3

[21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, June 2022. 1, 2

[22] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 3, 5, 7

[23] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 6

[24] Xiangteng He and Yuxin Peng. Fine-grained image classification via combining vision and language. In *CVPR*, 2017. 2

[25] Olivier J. Hénaff, Skanda Koppula, Jean-Baptiste Alayrac, Aaron van den Oord, Oriol Vinyals, and João Carreira. Efficient visual pretraining with contrastive detection. In *ICCV*, pages 10086–10096, October 2021. 2

[26] Zejiang Hou, Fei Sun, Yen-Kuang Chen, Yuan Xie, and Sun-Yuan Kung. Milan: Masked image pretraining on language assisted representation. *arXiv preprint arXiv:2208.06049*, 2022. 2

[27] Zhicheng Huang, Xiaojie Jin, Chengze Lu, Qibin Hou, Ming-Ming Cheng, Dongmei Fu, Xiaohui Shen, and Jiashi Feng. Contrastive masked autoencoders are stronger vision learners. *arXiv preprint arXiv:2207.13532*, 2022. 2

[28] Dat Huynh, Jason Kuen, Zhe Lin, Jiuxiang Gu, and Ehsan Elhamifar. Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7031, 2022. 5, 6

[29] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 2, 3, 5, 6, 8

[30] Armand Joulin, Laurens van der Maaten, Allan Jabri, and Nicolas Vasilache. Learning visual features from large weakly supervised data. In *ECCV*, 2016. 2

[31] Dahun Kim, Tsung-Yi Lin, Anelia Angelova, In So Kweon, and Weicheng Kuo. Learning open-world object proposals without learning to classify. *IEEE Robotics and Automation Letters*, 7(2):5453–5460, 2022. 4

[32] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models. *arXiv preprint arXiv:2209.15639*, 2022. 1, 3, 4, 7

[33] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022. 3

[34] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 12888–12900, 2022. 2

[35] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022. 3

[36] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. *arXiv preprint arXiv:2212.00794*, 2022. 2

[37] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *ECCV*, 2022. 4

[38] Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fahad Shahbaz Khan, Rao Muhammad Anwer, and Ming-Hsuan Yang. Class-agnostic object detection with multi-modal transformer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*, pages 512–531. Springer, 2022. 5

[39] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers. In *ECCV*, 2022. 1, 3, 5, 6

[40] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. 2014. 2

[41] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[42] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022. 2

[43] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Hanxiao Liu, Adams Wei Yu, Minh-Thang Luong, Mingxing Tan, and Quoc V. Le. Combined scaling for zero-shot transfer learning. *CoRR*, abs/2111.10050, 2021. 2, 8

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 3, 5, 6, 8

[45] Shafin Rahman, Salman Khan, and Nick Barnes. Improved visual-semantic alignment for zero-shot object detection. In *AAAI*, 2020. 3

[46] Hanoona Rasheed, Muhammad Maaz, Muhammad Uzair Khattak, Salman Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. *arXiv preprint arXiv:2207.03482*, 2022. 1, 2, 3, 5, 6

[47] Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. Learning visual representations with caption annotations. In *ECCV*, 2020. 2

[48] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 2

[49] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, 2019. 6

[50] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *CVPR*, pages 15638–15650, 2022. 6

[51] Weijie Su, Xizhou Zhu, Chenxin Tao, Lewei Lu, Bin Li, Gao Huang, Yu Qiao, Xiaogang Wang, Jie Zhou, and Jifeng Dai. Towards all-in-one pre-training via maximizing multi-modal mutual information. *arXiv preprint arXiv:2211.09807*, 2022. 2

[52] Josiah Wang, Katja Markert, Mark Everingham, et al. Learning models for object recognition from natural language descriptions. In *BMVC*, 2009. 2

[53] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022. 2

[54] Fangyun Wei, Yue Gao, Zhirong Wu, Han Hu, and Stephen Lin. Aligning pretraining for detection via object-level contrastive learning. In *NeurIPS*, 2021. 2

[55] Longhui Wei, Lingxi Xie, Wengang Zhou, Houqiang Li, and Qi Tian. Mvp: Multimodality-guided visual pre-training. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*, pages 337–353. Springer, 2022. 2

[56] Tete Xiao, Colorado J Reed, Xiaolong Wang, Kurt Keutzer, and Trevor Darrell. Region similarity representation learning. In *ICCV*, pages 10539–10548, October 2021. 2

[57] Shusheng Yang, Yixiao Ge, Kun Yi, Dian Li, Ying Shan, Xiaohu Qie, and Xinggang Wang. Masked visual reconstruction in language semantic space. *arXiv preprint arXiv:2301.06958*, 2023. 2

[58] 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. In *ICLR*, 2021. 6

[59] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *TMLR*, 2022. 5, 6

[60] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, November 2021. 6

[61] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 106–122. Springer, 2022. 5

[62] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *CVPR*, 2021. 1, 3, 4

[63] Bowen Zhang, Zhi Tian, Quan Tang, Xiangxiang Chu, Xiaolin Wei, Chunhua Shen, and Yifan Liu. Segvit: Semantic segmentation with plain vision transformers. *arXiv preprint arXiv:2210.05844*, 2022. 1

[64] Xinyu Zhang, Jiahui Chen, Junkun Yuan, Qiang Chen, Jian Wang, Xiaodi Wang, Shumin Han, Xiaokang Chen, Jimin Pi, Kun Yao, et al. Cae v2: Context autoencoder with clip target. *arXiv preprint arXiv:2211.09799*, 2022. 2

[65] Shiyu Zhao, Zhixing Zhang, Samuel Schulter, Long Zhao, Anastasis Stathopoulos, Manmohan Chandraker, Dimitris Metaxas, et al. Exploiting unlabeled data with vision and language models for object detection. In *ECCV*, 2022. 1, 2, 3, 5, 6

[66] Ye Zheng, Ruoran Huang, Chuanqi Han, Xi Huang, and Li Cui. Background learnable cascade for zero-shot object detection. In *ACCV*, 2020. 3

[67] Yiwu Zhong, Jing Shi, Jianwei Yang, Chenliang Xu, and Yin Li. Learning to generate scene graph from natural language supervision. In *ICCV*, 2021. 6

[68] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. Regionclip: Region-based language-image pretraining. In *CVPR*, 2022. 1, 2, 3, 4, 5, 6

[69] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *ECCV*, 2022. 3

[70] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 2

[71] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 1, 3, 5, 6

[72] Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama. Don't even look once: Synthesizing features for zero-shot detection. In *CVPR*, 2020. 3