Modulating CNN Features with Pre-Trained ViT Representations for Open-Vocabulary Object Detection

Xiangyu Gao, Yu Dai, Benliu Qiu, Hongliang Li
University of Electronic Science and Technology of China
{xygao, ydai, qbenliu}@std.uestc.edu.cn hlli@uestc.edu.cn

Abstract

Owing to large-scale image-text contrastive training, pre-trained vision language model (VLM) like CLIP shows superior open-vocabulary recognition ability. Most existing open-vocabulary object detectors attempt to utilize the pre-trained VLM to attain generative representation. F-ViT uses the pre-trained visual encoder as the backbone network and freezes it during training. However, the frozen backbone doesn't benefit from the labeled data to strengthen the representation. Therefore, we propose a novel two-branch backbone network design, named as ViT-Feature-Modulated Multi-Scale Convolutional network (VMCNet). VMCNet consists of a trainable convolutional branch, a frozen pre-trained ViT branch and a feature modulation module. The trainable CNN branch could be optimized with labeled data while the frozen pre-trained ViT branch could keep the representation ability derived from large-scale pre-training. Then, the proposed feature modulation module could modulate the multi-scale CNN features with the representations from ViT branch. With the proposed mixed structure, detector is more likely to discover novel categories. Evaluated on two popular benchmarks, our method boosts the detection performance on novel category and outperforms the baseline. On OV-COCO, the proposed method achieves 44.3 AP_{50}^{novel} with ViT-B/16 and 48.5 AP_{50}^{novel} with ViT-L/14. On OV-LVIS, VMCNet with ViT-B/16 and ViT-L/14 reaches 27.8 and 38.4 mAP_r.

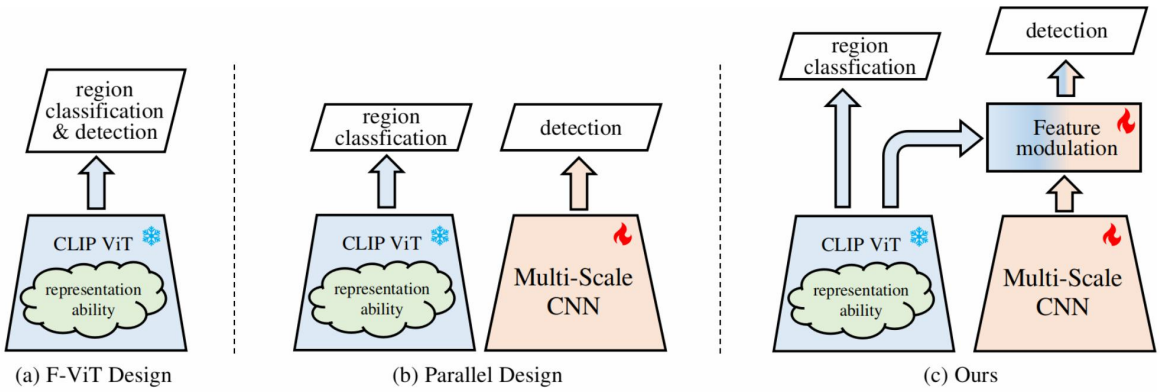
1. Introduction

With the advances in cross-modality learning, open-vocabulary object detection [32] (OVOD) is proposed. This task requires models to detect the novel targets beyond the training-category set via the language instruction. Compared to traditional closed-set detector, OVOD model improves the generalization ability from instance-level to category-level, which makes a further step in the real-world applications.

For most existing OVOD methods, pre-trained VLM is essential to implement open-vocabulary recognition. CLIP [18], a typical VLM pre-trained on an enormous number of image-text pairs, is able to encode image into the feature aligned with the text embedding space. To better satisfy the dense prediction task such as object detection, some works [7, 27, 36, 37] further improve CLIP, which transfer the visual encoder's representation ability from image-level to region-level. Based on the dense visual feature encoder, F-ViT [27] firstly builds open-vocabulary object detectors upon a frozen CLIP ViT backbone and achieves remark performance. In F-ViT, its backbone network is fine-tuned exclusively in a self-distillation manner at first. During the base training, F-ViT freezes its backbone to save the generalization ability. As shown in Figure 1(a), F-ViT uses the interpolated intermediate features from the frozen CLIP ViT for detection.

It is common knowledge that backbone network plays a key role in detection performance. In most detector designs, the backbone networks are optimized with the labeled bounding boxes during training, which ensures the detection performance. However, to save the generalization ability, the backbone network of F-ViT gives up exploiting the knowledge of labeled data during base training. Although its self-distillation strategy strengthens the local representation of CLIP ViT which may implicitly promotes detection, this pretext task treats the vision transformer more like a region classifier than a backbone network. Backbone network, as the feature extractor, needs to suit the downstream modules in a detector via the bounding box training. Thus, merely using self-distillation may not be adequate for feature extraction of detection.

A usual scheme in object detection is to initialize the backbone with the weights from pre-training and fine-tune it with training data. However, directly fine-tuning the pre-trained model with limited data will harm its open-vocabulary recognition ability, which degrades the model into a closed-set detector. As reported in [27], the performance on novel categories will degenerate when using trainable CLIP ViT. Besides, another straightforward



(a) F-ViT Design (b) Parallel Design (c) Ours

Figure 1. Different backbone paradigms for open-vocabulary object detection. (a) F-ViT uses a frozen CLIP ViT to extract image features for detection, which does not employ the base training data to strengthen the representation. (b) This paradigm use an extra trainable convolutional neural network, which is optimized with base training data for detection. However, the representation ability of CLIP ViT is not exploited. (c) Our design applies two-branch architecture, the representations from frozen CLIP ViT are utilized to modulate the features from trainable CNN. Thus, the final representations for detection benefits from both the pre-trained model and training data.

scheme is applying an extra trainable network to extract features parallelly for detection as shown in Figure 1(b). Whereas the additional network exploits the training data, such design isolates the CLIP ViT, which does not make use of the representation ability of VLM for detection. Hence, it is nature to ask a question: could a backbone structure utilize the information from both the base training data and the pre-trained VLM simultaneously?

Inspired by ViT-CoMer [30], we are motivated to design a two-branch backbone network, which could take advantages of the base training and pre-trained CLIP ViT. The proposed backbone network, named as ViT-Feature-Modulated Multi-Scale Convolutional network (VMCNet), consists of a vision transformer branch, a convolutional branch and a feature modulation module. Figure 1(c) shows the sketch of our design. Firstly, to retain the generalization ability, we also freeze the CLIP ViT but use it as a part of network. These ViT features are treated as intermediate products instead of the final output of backbone. Then, to exploit the base training, we use a trainable convolutional neural network (CNN) as the convolutional branch. This CNN branch has a simple structure and is able to produce multi-scale features. Finally, the proposed feature modulation module fuses the outputs from two branches to generate the final representation. As a result, VMCNet improves the detection performance on novel categories of OVOD detector.

In summary, our main contributions are listed as follow:

- We propose a novel backbone network for open-vocabulary object detection, named as **V**iT-Feature-**M**odulated Multi-Scale **C**onvolutional network (VMCNet). Via utilizing two-branch architecture, it leverages the strengths of different networks and exploits the information from both the base training data and the pre-trained VLM.
- We design a feature modulation module to merge the features from different branches. By this module, multi-scale features from trainable convolutional network are modulated with the adapted feature from frozen ViT. The process of feature fusion improves the quality of representation and boosts the detection performance on novel categories.
- We evaluate our proposed VMCNet on two popular benchmarks. The experimental results demonstrate that our method significantly improves the detect performance on novel categories. On OV-COCO, our method with ViT-B/16 achieves 44.3 AP_{50}^{novel}. When utilizing ViT-L/14, VMCNet could achieve 48.5AP_{50}^{novel}. Moreover, VMCNet with ViT-B/16 and ViT-L/14 attains 27.8 and 38.4 mAP_r on OV-LVIS, which is competitive with SOTA methods.

2. Related Work

Backbone Structure for Object Detection. In deep vision tasks, deep neural network was firstly applied in image classification. The advent of convolutional neural network (CNN) such as ResNet [9] strongly promoted the development of vision tasks. R-CNN [3] started to use the neural network pre-trained on classification dataset for feature extraction. From then on, employing backbone network in detector became a common practice. Besides CNN, transformer [22] architecture was introduced into visual field. In ViT [2], an image is divided into patches then encoded into a sequence. These attention-based designs [14, 24, 34] also

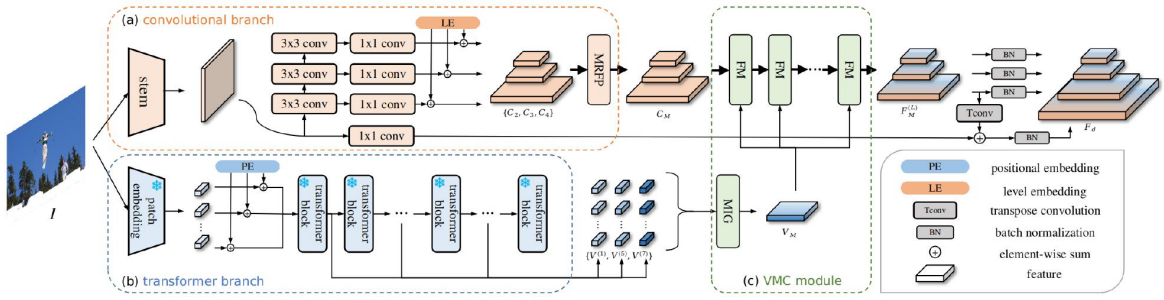


Figure 2. The overall architecture of proposed VMCNet. Flatten operation is omitted for clarity. Modules marked with snowflake are frozen, the others are optimizable during training. (a) Convolutional branch extracts multi-scale features from the input image. (b) Pre-trained transformer branch provides its intermediate features. (c) VMC module merges the outputs from two branches to generate final multi-scale features.

exhibited superior performances in vision tasks . Recently , [39] proposes a backbone network based on state space models (SSMs), which serves as a new option for detector design. To further exploit the strength of different architectures, some works [5, 15, 30] design the hybrid approaches which further boost the dense visual prediction tasks.

No matter whether the detector is CNN-based [8, 21] or query-based [1, 20], the detection prediction is derived from the extracted feature. Therefore, in traditional closed-set object detection, backbone networks are supposed to be optimized adequately during training to ensure the detection performance.

Inspired by ViT-CoMer [30], our design chooses the mixed-architecture which combines ViT and CNN. Distinct from these closed-set detection designs, not all modules in our backbone network are trainable. To satisfy the open-vocabulary setting, our backbone network attempts to exploit the information not only from the training data, but also the pre-trained VLM. Thus, a frozen CLIP ViT is employed as a branch of backbone network, and our feature fusion module performs unidirectional information injection instead of bidirectional interaction.

Backbone design for Open-Vocabulary Object Detection. One of the major issues in OVOD is how to alleviate over-fitting. So far, most OVOD designs still rely on bounding box training to enable the localization ability. However, this attribute inevitably leads to over-fitting on base categories. Another fact is that most OVOD detectors inherit the model structures from the closed-set designs. Backbone network, as the indispensable part of detector, faces the conflict between capturing localization ability and alleviate over-fitting on base-data.

Many works such as [4, 17, 23, 26, 35] choose to optimize the backbone networks with base data for detection, especially these based on the two-stage detectors like [8]. A common strategy is to replace the class-specific module into the class-agnostic and train the entire detector. Usually, the architecture of these backbones is CNN-based. With the additional improvement such as knowledge distillation [4, 23], pseudo labels [35] and etc, the detectors could perform well for OVOD task.

There are also some works [29, 37, 37] using the frozen backbone networks. Although ViT visual encoder outperforms the ResNet in image-level recognition, the CNN-based network can preserve region attribute better for dense prediction. With modification on the final layer like [37], CLIP ResNet could produce the dense feature for detection. Following two-stage detector structure, F-VLM [12] freezes its backbone and fine-tunes only the detector head during base training. CORA [29], the query-based method, also extracts image features with the frozen CLIP ResNet. To explore the potential of CLIP ViT on dense prediction task such as detection, CLIPSelf utilizes a self-distillation strategy, which enhances the recognition ability of CLIP ViT on region-level. Then, F-ViT uses this frozen CLIP ViT as the backbone network to realize the OVOD.

Unlike these single-branch backbone designs, our method utilizes the two-branch structure which consists of both the trainable branch and the frozen. Two types of features could be efficiently fused into a stronger representation.

3. Method

3.1. Preliminary

Our detector structure is based on F-ViT. Before detailing our backbone design, we will give a brief description of F-ViT to get a clear view of the entire OVOD method. Unless specified, we choose ViT-B/16 as the default setting for illustration.

The detection task could be divided into two sub-tasks, i.e., localization and classification. F-ViT follows the two-stage detector design [8]. The RoI features for classification and localization regression are derived from the intermedi-

ate features from the frozen ViT. We denote the output features of $i-$th transformer block in CLIP ViT as $V^{(i)}$, prediction of the detector could be represented as:

$$F_{RoI} = \phi(V^{(4)}, V^{(6)}, V^{(8)}, V^{(12)}) \quad (1)$$

$$b_p = localizer(F_{RoI}) \quad (2)$$

$$s_p = classifer(F_{RoI}) \quad (3)$$

where b_p is the predicted bounding box, s_p is the predicted category score for classification, and ϕ denotes the processing includes FPN, RPN, RoIAlign, etc.

Since the region-level representation of ViT is enhanced with self-distillation, F-ViT also applies the dense feature from last transformer block to improve the classification prediction. Specifically, they use RoI Align to attain the region feature V_{RoI} with RoI bounding box b_r. After normalization, the VLM score s_{VLM} is attained via calculating the cosine similarity between V_{RoI} and the encoded text feature T. The final classification score s is the weighted geometric average of s_p and s_{VLM}, where β and γ are hyperparameters.

$$V_{RoI} = RoIAlign(V^{(12)}, b_r) \quad (4)$$

$$s_{VLM} = softmax(\beta \cdot cos < V_{RoI}, T >) \quad (5)$$

$$s = s_p^{\gamma} \cdot s_{VLM}^{1-\gamma} \quad (6)$$

In our design, we focus on improving the quality of F_{RoI} by using the proposed backbone network. Except the backbone design, we follow the detector design of F-ViT to implement open-vocabulary object detection.

3.2. Overall Architecture

Our goal is to design a backbone which could efficiently utilize the information from both the base training data and pre-trained vision transformer. The overall architecture of VMCNet is illustrated in Figure 2, which mainly includes three parts: (a) Multi-Scale Convolutional Branch. (b) Vision Transformer Branch. (c) **V**iT-Feature-**M**odulated Multi-Scale **C**onvolutional(VMC) Module. These three components will be introduced specifically in Section 3.3, 3.4 and 3.5, respectively.

The input image I passes through both branches parallelly to obtain the multi-scale convolutional features and vision transformer features. Then, two types of features are fed into **V**it-**M**odulated **C**onvolutional (VMC) module, which modulates the multi-scale CNN features with the adapted ViT representation. The output of VMC module is then processed via batch normalization and transpose convolution to attain the final feature F_d. F_d could serve as the backbone features for downstream processing. Note that F_d contains the features with resolutions of 1/4, 1/8, 1/16 and 1/32, which is same as the setting in F-ViT.

3.3. Multi-Scale Convolutional Branch

Convolutional neural network is well-known for its properties of local continuity and multi-scale capabilities, besides, it usually have less computation burden than the attention-based network. Thus, we choose to utilize CNN as the optimizable branch to learn detection representation from base training. Inspired by ViT-CoMer [30], we utilize the early stage of its convolutional branch to extract multi-scale convolutional features. As shown in Figure 2, this light-weight version of CNN branch consists of a stem module [9], a stack of convolutional layers, layer embedding addition and a multi-receptive field feature pyramid (MRFP) module [30]. By using such structure, the additional computation cost is acceptable and we can customize the channel number of output.

Firstly, stem module processes an input image I with the shape of $H \times W \times 3$ into the feature with a resolution reduction of 1/4 of the original image. Then, this intermediate feature goes through three stride-2 3×3 convolutional layers to obtain the features with resolutions of 1/8, 1/16, and 1/32, respectively.

Next, feature at each scale is projected by its corresponding 1×1 convolutional layer. Except feature at the largest scale, projected features are added with layer embeddings. The output features are denoted as $C_1 \in R^{\frac{H}{4} \times \frac{W}{4} \times D}$, $C_2 \in R^{\frac{H}{8} \times \frac{W}{8} \times D}$, $C_3 \in R^{\frac{H}{16} \times \frac{W}{16} \times D}$, and $C_4 \in R^{\frac{H}{32} \times \frac{W}{32} \times D}$, respectively. For saving the computation cost, outputs excluding C_1 are fed into MRFP and VMC.

The structure of MRFP is specified in [30], this module could refine multi-scale features efficiently and expand receptive field. C_2, C_3 and C_4 are flattened and concatenated into C. We study the choice of MRFP number in Section 4.3.1. The output of convolutional branch is attained as follow, where $C, C_M \in R^{(\frac{HW}{8^2} + \frac{HW}{16^2} + \frac{HW}{32^2}) \times D}$:

$$C_M = MRFP(C) \quad (7)$$

3.4. Vision Transformer Branch

Pre-trained on an enormous number of image-text pairs, visual encoder in CLIP [18] could process an image into the encoded feature aligned with text feature space. CLIP-Self [27] further improved the visual transformer encoder with self-distillation for dense prediction task. Thus, we use the frozen visual encoder from [27] as ViT branch to attain the modulating feature.

In ViT branch, input image I firstly passes through a large kernel 16×16 convolutional layer (patch embedding). After added with the positional embedding, we attain the input feature $V^{(0)} \in R^{\frac{H}{16} \times \frac{W}{16} \times D}$. $V^{(0)}$ is fed into a series of vision transformer blocks to attain its intermediate features.

To make full use of the representation ability of the frozen ViT, we collect multiple intermediate features from

this branch. Under ViT-B/16 setting, we choose $V^{(1)}$, $V^{(5)}$ and $V^{(7)}$ for feature modulation. Study about the choice of ViT layers is presented in Section 4.3.2. Since not all transformer blocks are used during training phase, we could skip the computation of remaining blocks during training. In inference phase, we apply the whole vision transformer to generate feature map from the last block for dense prediction.

3.5. VMC Module Design

Inspired by ViT-CoMer [30], a strong backbone network utilizing the pretrained ViT via feature interaction, we design the VMC module referring to its CNN-Transformer bidirectional fusion interaction(CTI) block. As shown in Figure 2, VMC module mainly consists of a modulating information generation (MIG) block and L feature modulation (FM) blocks. Notice that there are several crucial differences between ViT-CoMer and our design.

First, VMC module focuses on injecting the information from a sophisticated visual encoder into CNN features, which is unidirectional instead of bidirectional. To keep the rich knowledge in visual encoder, the vision transformer in VMC is frozen while ViT-CoMer is fine-tunable. Moreover, VMC module is applied merely at the end of two branches, which owns different structure from it. Our method is designed for OVOD task, while ViT-CoMer aims to enhance the closed-set object detector.

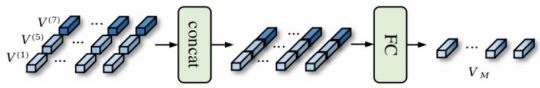


Figure 3. Structure of modulating information generation module. FC denotes the fully connected layer.

Modulating Information Generation. We attempt to utilize the ViT features collected from different blocks to generate a unified representation. Thus, we propose a novel module named as modulating information generation. Its structure is illustrated in Figure 3. Firstly, the ViT features $V^{(1)}$, $V^{(5)}$ and $V^{(7)}$ are concatenated along the channel dimension. This concatenated high-dimensional features contains representation from different depth in ViT. Then, a linear projection layer is applied, which not only reduce the channel number from $3 * D$ to D, but also simply adapting the ViT features to the downstream processing. We implement this linear projection with a fully connected layer. Via MIG, we could attain the adaptive modulating feature. The processing of MIG could be denoted as:

$$V_M = FC(concat(V^{(1)}, V^{(5)}, V^{(7)})) \tag{8}$$

Feature Modulation. With the adaptive modulating features from MIG and the multi-scale features from CNN

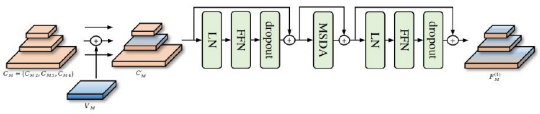


Figure 4. Structure of feature modulation block. We illustrate processing in the first block as example.

branch, we achieve the feature modulation by stacking a serial of feature modulation (FM) blocks. As shown in Figure 2, FM block takes the multi-scale features and the modulating feature as inputs. The output of the last block is sent to the next in a manner of cascaded refinement. All blocks use the same modulating feature V_M.

FM shares the same structure as ViT-to-CNN (VtoC) block in ViT-CoMer, its structure is illustrated in Figure 4. This block consists of a Multi-scale deformable attention (MSDA), two layer normalization operations, two feed-forward networks and three skip connection sums.

In first FM block, V_M is added into the CNN feature at 1/16 scale. We denote the new multi-scale features as C'_M. Then C'_M goes through a layer normalization layer, a feed-forward network and a dropout layer, the processing is represented as below:

$$C'_M = \{C_{M2}, C_{M3} + V_M, C_{M4}\} \tag{9}$$
$$C_F = Dropout(FFN(LN(C'_M))) + C'_M \tag{10}$$

Whereas the element-wise sum operation simply fuses the features, the representation from ViT is merely injected into the features at second scale. As we know, deformable attention (DA) utilizes the reference points to sample the features to attain the values. In MSDA, each scale will provide a fixed number of reference points, that is, ViT representation could be scattered into the features at all scales via MSDA. The output of MSDA is then processed similarly as the above:

$$C'_F = MSDA(C_F) + C_F \tag{11}$$
$$F_M^{(1)} = Dropout(FFN(LN(C'_F))) + C'_F \tag{12}$$

Through the operations in first FM, features at each scale of $F_M^{(1)}$ has carried the representations from frozen ViT. In the next FM block, we repeat the processing as illustrated above to further inject the ViT representations into the multi-scale features. For i-th block, feature modulation is denoted as:

$$F_M^{(i)} = FM(F_M^{(i-1)}, V_M) \tag{13}$$

We also study the choice of L in Section 4.3.1. The output of final block is denoted as $F_M^{(L)}$, we interpolate the

feature with largest scale in $F_M^{(L)}$ and add it with C_1. After batch normalization, this four-scale features could be used as the backbone features for downstream processing.

$$F_{d1} = BN(Tconv(F_{M2}^{(L)}) + C_1) \tag{14}$$

$$F_{di} = BN(F_{Mi}^{(L)}), i = 2, 3, 4 \tag{15}$$

$$F_d = \{F_{d1}, F_{d2}, F_{d3}, F_{d4}\} \tag{16}$$

4. Experiments

4.1. Experiment Setup

4.1.1 Datasets and evaluation metrics

Our method is evaluated on two popular open-vocabulary object detection datasets, OV-COCO [13] and OV-LVIS [6].

OV-COCO. MSCOCO is one of the most commonly used datasets in object detection. To evaluate a detector under open-vocabulary setting, OVR-CNN [32] proposes to build the base set with 48 categories of its data, and the other 17 categories of data are used as the novel. We follow this rule to train our method with the base set data and evaluate it with the set including both the base and the novel. As the behaviors of existing works, bounding box AP at IoU threshold 0.5 of novel categories is employed as the main metric, which denoted as AP_{50}^{novel}.

OV-LVIS. LVIS is a large-scale dataset for instance segmentation, which contains 1203 categories of objects. The categories is split into rare (337 categories), common (461) and frequent(405) based on the number of images the they appear. We follow ViLD [4] to split the dataset, rare categories are split into the novel while common and frequent categories are split into the base. The mask AP averaged on IoUs from 0.5 to 0.95 of rare categories, denoted as mAP_r, is chosen as the main metric for evaluation.

4.1.2 Implementation Details

Our detection framework is based on F-ViT [27]. We use the self-distilled visual encoder from CLIPSelf to initialize the weights of vision transformer branch. For OV-COCO, we use the CLIPSelf with proposal distillation setting while OV-LVIS is the patch distillation setting. We conduct the experiments with 8 RTX 3090 GPUs and set the batch size as 1 on each GPU. For fair comparisons, our method inherits the configurations of baseline except the proposed part. Following the implementation of F-ViT, we use the same training scheme. Utilizing AdamW optimizer with a learning rate of 1.25×10^{-5}, detector for OV-COCO is trained for 3 epochs while the one for OV-LVIS is 48-epoch training.

4.2. Quantitative Comparisons

OV-COCO. We report the comparison with previous methods in Table 1. VMCNet based on ViT-B/16 is denoted as VMCNet-B while VMCNet-L refers to the one based on ViT-L/14. To get rid of the influence from experiment environment, we run the baseline method under our local environment and display the results. Under ViT-B/16 setting, our method achieves 44.3 AP_{50}^{novel}, which is even comparable to the baseline with ViT-L/14. The performance can be further improved when increasing the scale of network. F-ViT equipped with VMCNet-L achieves 48.5 AP_{50}^{novel} and also outperforms the baseline method by an obvious margin of 4.0%. Through the evaluation on OV-COCO, we found VMCNet could effectively boost the detection performances on novel categories.

Method	Backbone	AP_{50}^{novel}(%)
ViLD [4]	RN50	27.6
Detic [38]	RN50	27.8
F-VLM [12]	RN50	28.0
OV-DETR [31]	RN50	29.4
BARON-KD [26]	RN50	34.0
CLIM BARON [28]	RN50	36.9
SAS-Det [35]	RN50-C4	37.4
EdaDet [19]	RN50	37.8
SIA [25]	RN50x4	41.9
CORA+ [29]	RN50x4	43.1
RO-ViT [11]	ViT-L/16	33.0
CFM-ViT [10]	ViT-L/16	34.1
F-ViT+CLIPSelf [27]	ViT-B/16	37.6
BIND [33]	ViT-L/16	41.5
F-ViT+CLIPSelf [27]	ViT-L/14	44.3
F-ViT+CLIPSelf † [27]	ViT-B/16	38.7
F-ViT+VMCNet	VMCNet-B	44.3 (+5.6)
F-ViT+CLIPSelf † [27]	ViT-L/14	44.5
F-ViT+VMCNet	VMCNet-L	**48.5** (+4.0)

Table 1. Comparison with state-of-the-art methods on OV-COCO benchmark. '†' denotes that this result is obtained from the reimplemented experiments under our local environment.

OV-LVIS. Table 2 lists the state-of-the-art methods on OV-LVIS. VMCNet also surpasses previous state of the arts. Compared to the baseline with ViT-B/16, our method can bring +2.0 mAP_r gain. Under ViT-L/14 setting, VMCNet can still outperform the baseline approach by 1.5 mAP_r. Therefore,VMCNet is proved to be effective on detecting objects of novel categories.

4.3. Ablation Experiments

We conduct ablation studies on OV-COCO benchmark under ViT-B/16 setting. Unless specified, we report performances with the highest AP_{50}^{novel}, the default setting is the same as the statements in Section 3.

Method	Backbone	mAP$_r$(%)
ViLD [4]	RN50	16.6
OV-DETR [31]	RN50	17.4
BARON-KD [26]	RN50	22.6
EdaDet [19]	RN50	23.7
CORA+ [29]	RN50x4	28.1
SASDet [29]	RN50x4-C4	29.1
F-VLM [12]	RN50x64	32.8
OWL-ViT [16]	ViT-L/14	25.6
RO-ViT [11]	ViT-L/16	32.4
BIND [33]	ViT-L/16	32.5
CFM-ViT [10]	ViT-L/16	33.9
RO-ViT [11]	ViT-H/16	34.1
F-ViT+CLIPSelf	ViT-B/16	25.8
F-ViT+VMCNet	VMCNet-B	**27.8**(+2.0)
F-ViT+CLIPSelf	ViT-L/14	36.9
F-ViT+VMCNet	VMCNet-L	**38.4**(+1.5)

Table 2. Comparison with state-of-the-art methods on OV-LVIS.

ViT	CNN	FM*	FM	MIG	AP_{50}^{novel}(%)	AP_{50}^{base}(%)	TParams
✓	×	×	×	×	38.70	53.94	21.02M
✓	✓	×	×	×	18.40	13.89	18.50M
✓	✓	✓	×	×	23.06	15.01	24.86M
✓	✓	×	✓	✓	**44.33**	49.93	24.86M

Table 3. Ablation study on main components of VMCNet. 'FM*' refers to the FM block **without** V_M feature addition. 'TParams' means the number of trainable parameters in the whole detector.

4.3.1 Effectiveness of model structure

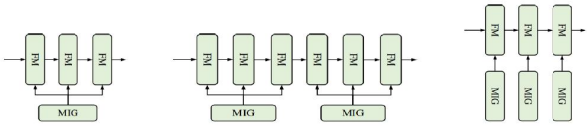
In this part, we attempt to analyze the effectiveness of components in VMCNet and evaluate the feasibility of these structures.

The ablation study on the main components is shown in Table 3. We start from the baseline method which only applies the CLIP ViT, its performance is reported in first row. In second row, the last few transpose convolutional layers to interpolate the ViT features are removed and backbone is replaced by the trainable CNN, where ViT only serves as the RoI classifier. This design corresponds to the illustration in Figure 1(b), where detector could only rely on the features extracted by the simple CNN and thus have poor performance. In third row, we further add the abridged FM blocks which does not add V_M, though the performance is improved by a little, detector still suffers from lack of ViT representation. In the bottom line, the entire VMCNet is implemented. We can see that the increase on parameter cost is lower than 4M. Therefore, these main components can effectively utilize the information from both the frozen ViT and base training data.

Besides, we explore the strategy of placing FM in VMC modules. As shown in Table 4, we presents 5 strategies. We can see that the highest AP_{50}^{novel} is achieved in the second row, that is, the optimal value of L is 3. To better understand these strategies, the local designs in 2-nd, 4-th and 5-th row are illustrated in Figure 5(a), (b) and (c), respectively. Compared to the 2-nd row, the 4-th adds a repeated structure behind, which does not improve performance. In 5-th line, each FM block is supported by a different MIG module, its result shows using more MIG modules can not bring gain, either. Therefore, we choose the strategy in second row for OV-COCO setting.

L	AP_{50}^{novel}(%)	AP_{50}^{base}(%)
1×2	43.11	49.42
1×3	**44.33**	49.93
1×4	43.95	50.12
2×3	43.65	50.47
3×1	43.85	50.30

Table 4. Ablation study on the strategy of placing FM blocks.'$a \times b$' represents that there are a groups of FMs, each group is supported by a MIG module and contains b FM blocks.



(a) 1×3 (b) 2×3 (c) 3×1

Figure 5. Optional strategies of placing FM blocks in VMC module.

N_{MRFP}	AP_{50}^{novel}(%)	AP_{50}^{base}(%)
0	42.93	49.71
1	**44.33**	49.93
2	42.99	49.55

Table 5. Effectiveness of N_{MRFP}. 'N_{MRFP}' denotes the number of MRFP modules.

We also analyze the impact of the number of MRFP modules in CNN branch. The first row in Table 5 corresponds to the structure removing MRFP from VMCNet, which leads a drop of 1.4 AP_{50}^{novel}. The bottom line is the result of cascading two MRFP modules, which shows that increasing N_{MRFP} does not improve the performance. Thus, only one MRFP module is used in our CNN branch design.

4.3.2 Effectiveness of ViT branch

In this part, we conduct ablation experiments on the used ViT feature layers as shown in Table 6. The default setting we use is the 3-rd column. In first three rows, one of the de-

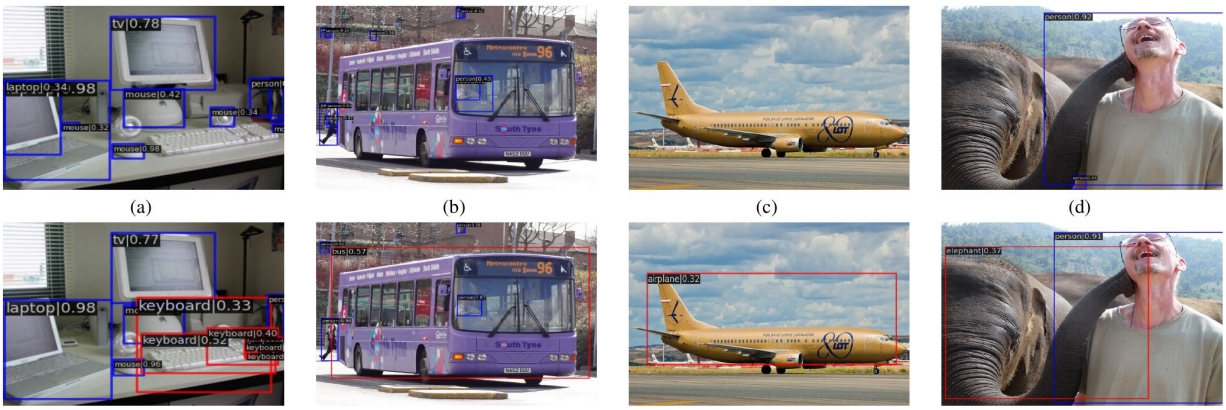


Figure 6. Comparison of visual results. The detection results from baseline are displayed in the upper line while ours are in the bottom line. Both the detector architectures are based on ViT-B/16. We set the score threshold as 0.3 to filter the predicted bounding box whose classification are lower than the value. Blue bounding box represents that the predicted category belongs to the base while red box belongs to the novel.

vit layer ids	$AP_{50}^{novel}(\%)$	$AP_{50}^{base}(\%)$
1,5	39.73	42.29
1,7	43.25	49.16
5,7	43.47	49.93
1,5,7	44.33	49.93
1,5,7,9	36.02	48.61
6,10,12	33.59	49.47

Table 6. Effectiveness of the used ViT feature layers.

fault feature layers is removed. As a result, these results are inferior to the default setting in the 4-th line. However, in fifth row, the detection performance on novel categories degenerates obviously when the extra ViT layer $V^{(9)}$ is added. In the last row, features from the deep blocks of ViT are applied, which leads to the poorer performance. Thus, we find that the ability to detect novel categories may benefit from features of some certain layers of frozen ViT. In OV-COCO setting, features from the high level of ViT may even hurt the performances.

At last, there still remains a doubt whether the gain on performance is due to the transformer model structure or the knowledge of pre-training. To find out the answer, we unfreeze the ViT branch and use an extra frozen ViT for RoI classification. The results are presented in Table 7. The second row shows that using the trainable ViT branch leads to more parameter cost and does not enhance the performance. Therefore, the proposed method can utilize the knowledge of pre-training to improve the detection performances.

ViT state	$AP_{50}^{novel}(\%)$	$AP_{50}^{base}(\%)$	TParams
frozen	**44.33**	49.93	24.86M
trainable	38.75	47.59	75.65M

Table 7. Frozen ViT branch vs trainable ViT branch.

4.4. Qualitative Visualization

In Figure 6, we visualize the detection results of the baseline method and ours. These image samples are collected from COCO validation set. From the visual comparisons, we observe that our method is more confident to recognize the novel objects. For example, in second column, the baseline fails to detect the bus in image, while ours could predict it with a relatively high confidence score.

5. Conclusion

In this paper, we propose a two-branch backbone network for OVOD task, named as VMCNet. Our method combines the advantages of CNN and ViT. The frozen ViT branch saves the generalization ability while the trainable CNN learns information of the base training data. At last, the proposed VMC module could modulate multi-scale convolutional features with the representations from ViT branch. Our design provides an effective scheme to utilize the knowledge from both the pre-training and base training. As a result, VMCNet improves the detection performances on novel categories effectively and efficiently. Experimental results demonstrate that our backbone network outperforms the baseline and reaches the level of state-of-the-art methods.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 213–229. Springer, 2020. 3

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 2

[3] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 580–587. IEEE Computer Society, 2014. 2

[4] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 3, 6, 7

[5] Jianyuan Guo, Kai Han, Han Wu, Yehui Tang, Xinghao Chen, Yunhe Wang, and Chang Xu. CMT: convolutional neural networks meet vision transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 12165–12175. IEEE, 2022. 3

[6] Agrim Gupta, Piotr Dollár, and Ross B. Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5356–5364. Computer Vision Foundation / IEEE, 2019. 6

[7] Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. *CoRR*, abs/2404.08181, 2024. 1

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2980–2988. IEEE Computer Society, 2017. 3

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 2, 4

[10] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Contrastive feature masking open-vocabulary vision transformer. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 15556–15566. IEEE, 2023. 6, 7

[11] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Region-aware pretraining for open-vocabulary object detection with vision transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 11144–11154. IEEE, 2023. 6, 7

[12] Weicheng Kuo, Yin Cui, Xiuye Gu, A. J. Piergiovanni, and Anelia Angelova. Open-vocabulary object detection upon frozen vision and language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 3, 6, 7

[13] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014. 6

[14] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9992–10002. IEEE, 2021. 2

[15] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 11966–11976. IEEE, 2022. 3

[16] Matthias Minderer, Alexey A. 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. *CoRR*, abs/2205.06230, 2022. 7

[17] Chau Pham, Truong Vu, and Khoi Nguyen. LP-OVOD: open-vocabulary object detection by linear probing. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024*, pages 768–777. IEEE, 2024. 3

[18] 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 Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. 1, 4

[19] Cheng Shi and Sibei Yang. Edadet: Open-vocabulary object detection using early dense alignment. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 15678–15688. IEEE, 2023. 6, 7

[20] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, and Ping Luo. Sparse R-CNN: end-to-end object detection with learnable proposals. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 14454–14463. Computer Vision Foundation / IEEE, 2021. 3

[21] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: fully convolutional one-stage object detection. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 9626–9635. IEEE, 2019. 3

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. 2

[23] Luting Wang, Yi Liu, Penghui Du, Zihan Ding, Yue Liao, Qiaosong Qi, Biaolong Chen, and Si Liu. Object-aware distillation pyramid for open-vocabulary object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 11186–11196. IEEE, 2023. 3

[24] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 548–558. IEEE, 2021. 2

[25] Zishuo Wang, Wenhao Zhou, Jinglin Xu, and Yuxin Peng. SIA-OVD: shape-invariant adapter for bridging the image-region gap in open-vocabulary detection. In Jianfei Cai, Mohan S. Kankanhalli, Balakrishnan Prabhakaran, Susanne Boll, Ramanathan Subramanian, Liang Zheng, Vivek K. Singh, Pablo César, Lexing Xie, and Dong Xu, editors, *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pages 4986–4994. ACM, 2024. 6

[26] Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 15254–15264. IEEE, 2023. 3, 6, 7

[27] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 1, 4, 6

[28] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Wentao Liu, and Chen Change Loy. CLIM: contrastive language-image mosaic for region representation. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 6117–6125. AAAI Press, 2024. 6

[29] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. CORA: adapting CLIP for open-vocabulary detection with region prompting and anchor pre-matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 7031–7040. IEEE, 2023. 3, 6, 7

[30] Chunlong Xia, Xinliang Wang, Feng Lv, Xin Hao, and Yifeng Shi. Vit-comer: Vision transformer with convolutional multi-scale feature interaction for dense predictions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 5493–5502. IEEE, 2024. 2, 3, 4, 5

[31] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary DETR with conditional matching. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IX*, volume 13669 of *Lecture Notes in Computer Science*, pages 106–122. Springer, 2022. 6, 7

[32] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 14393–14402. Computer Vision Foundation / IEEE, 2021. 1, 6

[33] Heng Zhang, Qiuyu Zhao, Linyu Zheng, Hao Zeng, Zhiwei Ge, Tianhao Li, and Sulong Xu. Exploring region-word alignment in built-in detector for open-vocabulary object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 16975–16984. IEEE, 2024. 6, 7

[34] Qinglong Zhang and Yu-Bin Yang. Rest: An efficient transformer for visual recognition. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 15475–15485, 2021. 2

[35] Shiyu Zhao, Samuel Schulter, Long Zhao, Zhixing Zhang, B. G. Vijay Kumar, Yumin Suh, Manmohan Chandraker, and Dimitris N. Metaxas. Taming self-training for open-vocabulary object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13938–13947. IEEE, 2024. 3, 6

[36] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition,*

CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, pages 16772–16782. IEEE, 2022. 1

[37] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from CLIP. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXVIII*, volume 13688 of *Lecture Notes in Computer Science*, pages 696–712. Springer, 2022. 1, 3

[38] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IX*, volume 13669 of *Lecture Notes in Computer Science*, pages 350–368. Springer, 2022. 6

[39] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. 3