Pattern Recognition 171 (2026) 112187

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

GridCLIP: One-stage object detection by grid-level CLIP representation learning

Jiayi Lin, Shitong Sun*, Shaogang Gong

Queen Mary University of London, London, E1 4NS, United Kingdom

ARTICLE INFO

Keywords:
Object detection
Open-vocabulary learning
Long-tail learning

ABSTRACT

CLIP provides a shared image-text representation space with rich and diverse vocabulary, enabling object detection in undersampled and unseen categories. Recent CLIP-based object detection works show two-stage detectors typically outperform one-stage designs, but with significantly higher computational costs. A fundamental limitation of a two-stage detector is region-level alignment (distillation), which requires hundreds of image encoder forward passes from both the detector and CLIP in each image. In this work, we propose GridCLIP, a one-stage detector that requires only a single image encoder inference per input image, achieving up to 43× faster training and 5× faster inference compared to its two-stage counterpart ViLD, while substantially narrowing the accuracy gap. GridCLIP introduces a dual alignment strategy to learn fine-grained, grid-level representations: (1) grid-level alignment: learning grid-level features aligned with CLIP text encoder using annotated category labels, and (2) image-level alignment: aggregating grid-level features into an image-level representation aligned with the CLIP image encoder, which allows GridCLIP to learn grid-level representations of a broad range of categories, especially undersampled and unseen categories. Experiments on the LVIS benchmark show that GridCLIP achieves competitive results, with strong generalization to COCO and VOC, demonstrating its efficiency and effectiveness as a CLIP-based detector.

1. Introduction

Simultaneous multi-category object detection aims to both recognize (classify) and detect (locate) all instances of the given categories in an image. A significant challenge in training a good detector is the cost of labeling a large-scale dataset on a broad range of object categories with balanced data distributions. Existing detection datasets are often imbalanced with a long-tail distribution across categories [1] where some object categories have only a few or zero training sample(s). To address the challenges of both undersampled and unseen categories, recent approaches [2,3] have leveraged CLIP [4] by exploiting its joint image-text representation space to associate visual inputs with a wide range of categories.

These CLIP-based object detectors fall into two main categories: two-stage [2,5–9] and one-stage [10–12] detectors. While two-stage detectors generally outperform their one-stage counterparts, they incur significantly higher computational costs, primarily due to region-level alignment that requires hundreds of forward passes through both the detector and CLIP's image encoder per input image. This alignment, however, enables direct supervision from CLIP's image encoder, which is especially beneficial for undersampled and unseen categories (hereafter referred collectively as "minority categories"). One-stage detectors, by contrast, avoid region proposals and typically align with CLIP's text embeddings, trading off accuracy for efficiency. Methods like HierKD [12] and YOLO-World [13] attempt to bridge this gap through global or region-text alignment, but they often rely on large-scale image-caption datasets and pseudo-labels. More importantly, only aligning to CLIP text embedding space can introduce a fundamental misalignment between visual and textual spaces, as text embeddings derived from image or region descriptions may be suboptimal compared to features extracted directly from CLIP's image encoder. The key challenge is thus to design efficient one-stage architectures that can leverage the rich visual cues of CLIP's image encoder without incurring prohibitive computation or relying on costly textual supervision–offering a more grounded, scalable approach to learning minority categories.

To address the aforementioned limitations, we propose a one-stage open-vocabulary detector, GridCLIP, which leverages direct alignment with CLIP's image encoder to enable more grounded and efficient generalization to minority categories. GridCLIP introduces a dual alignment strategy comprising grid-level and image-level alignment. The *grid-level alignment* focuses on learning fine-grained spatial features (i.e., feature map pixels) by aligning them with CLIP's text embeddings, supervised using a detection dataset with limited annotated categories. To incorporate knowledge from CLIP's image encoder and enhance learning of

* Corresponding author.
E-mail addresses: jiayi.lin@qmul.ac.uk (J. Lin), shitong.sun@qmul.ac.uk (S. Sun), s.gong@qmul.ac.uk (S. Gong).

https://doi.org/10.1016/j.patcog.2025.112187
Received 18 December 2024; Received in revised form 22 June 2025; Accepted 20 July 2025
Available online 29 July 2025

minority categories, the grid-level features are further aggregated into an image-level representation and aligned with the global embedding produced by a fixed CLIP image encoder, which serves as a teacher. This process, referred to as *image-level alignment*, facilitates effective visual-to-visual supervision without relying on text-based adaptation. Crucially, since the backbone network is shared across both grid-level and image-level pathways, the grid-level embeddings implicitly inherit knowledge from both CLIP's text and image encoders. Together, these strategies allow GridCLIP to retain the efficiency of a one-stage detection pipeline while significantly improving its ability to generalize in open-vocabulary settings.

Overall, we propose a one-stage detector GridCLIP, which exploits CLIP to supplement the knowledge of minority categories in downstream detection datasets by simultaneously applying grid-level and image-level alignments, narrowing the performance gap from typical two-stage detectors, while requiring a much shorter training time (43 times less compared to ViLD [2]) and test time (5 times faster), without the need of extra image-caption datasets and pseudo labeling. Our contributions are:

- We exploit CLIP to supplement the missing knowledge of undersampled and unseen categories in training a one-stage detector, mitigating the poor performance due to the long-tail data distribution in most existing detection training data.
- We propose a simple yet effective visual-to-visual knowledge distillation method for learning undersampled and unseen categories in constructing a one-stage CLIP-based detector, providing 2.4 AP gains on unseen categories compared to the baseline.
- Without using extra pretraining processes or additional fine-tuning datasets, GridCLIP is capable of handling Open-Vocabulary Object Detection with considerable scalability and generalizability, reaching the comparable performance to two-stage detectors with much higher training and inference speed.

2. Related works

Vision-language pretrained model (VLM). Visual-only pretrained models had long dominated the pretraining paradigm until the emergence of Vision-Language Models (VLMs), which incorporate natural language as supervision. VLMs enhance the model's generalization capability by learning to align visual concepts with language expressions, often beyond manually predefined categories. Recently, a significant number of VLMs have been proposed [4,14], leveraging large-scale image-text datasets in an unsupervised or weakly supervised manner. These models usually have both image and text encoders to generate corresponding features that can be aligned in a cross-modality representational space for corresponding image-text matching. Utilizing these alignment spaces helps zero-shot transfer to a wide range of downstream visual recognition tasks, such as object detection [6], segmentation [15], classification [16–18], and video moment retrieval [19]. As one widely-used instance, CLIP [4] is trained on 400 million image-text pairs, which extends significantly the generalizability and usability of the learned image embeddings to align to broad categories. CLIP is widely applied both in downstream tasks oriented pretraining [5,20] and in fine-tuning for downstream tasks [2,11].

Object detection using VLM. OVR-CNN [21] was the first to use natural language (captions) for object detection. Recent detectors leverage large-scale image-text datasets to learn generalizable representations, but some VLMs, like GLIP [20,22] and DetCLIP series [23], require costly large-scale annotation datasets. In contrast, we focus on unsupervised VLMs like CLIP, using small annotation datasets to transfer knowledge to broader categories (Fig. 1).

To learn knowledge of seen (base) categories, most detectors [2,3] replace detection head classifiers with VLM text embeddings. Recent detectors mainly improve the learning in two aspects: learning better text embeddings of categories and extracting image embeddings to align with these text embeddings. (1) For generating better text embeddings, also called *Prompt Learning*, current approaches can be roughly classified as template-based and learnable prompting. The template-based one uses fixed incomplete sentences that can accept labels to build complete sentences [2,5,12], while the learnable prompting methods concatenate learnable parameters with category labels as the input, where the prompt is implicitly learned during fine-tuning [6,7,24]. GridCLIP uses template-based prompting as in the original CLIP and current OVOD detectors [2,12] for simplicity and scalability. (2) For extracting image embeddings to align with text embeddings, two-stage

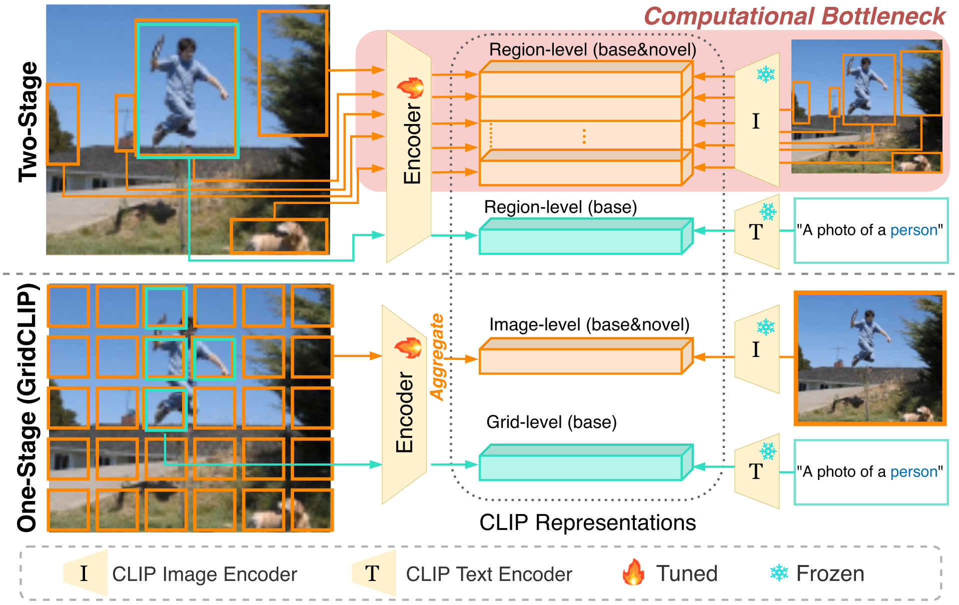


Fig. 1. Two-stage detectors (top) [2,6] perform region-level alignment to distill knowledge from CLIP image encoder, suffering from huge computational bottleneck. Specifically, they requires a large number of proposals for each image during distillation, which requires inference of CLIP image encoder and the detector's encoder for multiple times (up to 1000 in ViLD [2]). In contrast, our one-stage detector GridCLIP (bottom) aligns multiple grid-level representations that can be captured in a single pass of the encoder due to the nature of convolutional networks. These grid-level representations are then aggregated into an image-level representation, enabling efficient alignment with the CLIP image encoder in image level.

detectors [2,6,25] use the embedding of cropped object bounding-box proposals. However, they need to train a region proposal network first and require multiple inferences of the CLIP image encoder to compute the visual embedding for each region, which is relatively inefficient. In comparison, a one-stage detector [11,26,27] aligns parts in an image represented by grid-level embeddings. DenseCLIP [11] adapts it as aligning grid-level image features with text, while it can only perform under closed-set settings. Our method uses grid-level alignment for efficiency while preserving CLIP's original alignment space for better generalization, without requiring extra datasets.

To learn knowledge of unseen (novel) categories, several approaches [3,7,28] use extra knowledge from external datasets extra image-caption or labeled datasets like CC3M [29] or ImageNet-21K [30], making the training process complicated and resource-consuming. HierKD [12] introduces the global stage distillation method, which learns the image representation from the text features of image captions. Recently, YOLO-World [13] requires a pre-training process on 32 NVIDIA V100 GPUs on over a million of annotated images. However, we argue that CLIP has been trained over a broad vocabulary and has the ability to provide visual embeddings of various categories. Therefore, we explore the original CLIP representation space to learn unseen categories by applying knowledge distillation on CLIP in a more spatially fine-grained level.

In summary, existing object detection approaches face two critical limitations: 1) Two-stage detectors extract proposal-level features requires multiple CLIP passes (computational bottleneck). 2) Existing one-stage methods (e.g., HierKD) mostly use CLIP text encoder with image captions as proxy supervision. This however creates misalignment between caption semantics and actual visual content and further require extra image-text pairs and contrastive learning with large batch sizes and extensive resources. GridCLIP addresses these limitations by directly aligning grid-level features with CLIP embeddings in a one-stage detector framework without requiring image captions or multiple CLIP passes.

3. Approach

The overall model design of GridCLIP is shown in Fig. 2. We first introduce the strategy of adopting CLIP embeddings for the detection task, and then present the approach for simultaneously mapping the CLIP representation by both grid-level and image-level alignments based on a one-stage detector FCOS [27].

3.1. Adapting CLIP for detection

CLIP consists of an image encoder (ResNet [31] or ViT [32]) and a text encoder (Transformer [33]), which together form the alignment space of visual and language embeddings. However, CLIP's image embedding represents an entire image as a single abstract feature vector without spatial information, while its text embedding is designed for sentences rather than individual category labels used in detection. Therefore, further adaptation for the detection task is necessary.

Generating image embedding. The original CLIP image feature $\overline{z}$ is a single high-dimensional feature vector representing an entire image without spatial information. To get the grid-level feature z , inspired by DenseCLIP [11], we use the other feature from the last layer of the CLIP image encoder. Specifically, taking the ResNet50 encoder as an example, the final output feature in the 5-th stage $C_5 \in \mathbf{R}^{H_5 \times W_5 \times D_5}$ first undergoes global average pooling to get the image-level feature $\overline{C}_5 \in \mathbf{R}^{1 \times 1 \times D_5}$, where H_5, W_5, D_5 are the height, width and number of channels of the feature in the 5-th stage of the ResNet50. Then the concatenated features $\left[\overline{C}_5, C_5\right]$ are fed into a Multi-Head Self-Attention (MHSA) layer [33] as follows,

$$[\overline{z}, z] = \mathrm{MHSA}\left(\left[\overline{C}_5, C_5\right]\right). \tag{1}$$

In CLIP, the output with spatial information z is discarded and $\overline{z}$ is used to match with the text embedding. However, as illustrated in DenseCLIP, the symmetric nature of MHSA enables grid-level features to mirror image-level features, leading to alignment with text embeddings. Therefore, we adopt both z and $\overline{z}$ to generate our grid-level and image-level embeddings respectively with a few adaptation layers.

From label to text embedding. In CLIP, to create a dataset classifier from label text, a set of template-based prompts like "a photo of a {*object*}." are applied, where *object* is any of the target category names. Then the multiple prompts for a single label are aggregated. Although there are several learnable prompting methods for CLIP [24] that fine-tune with the downstream tasks, we follow the template-based one for simplicity and scalability. We use a template-based variant from ViLD [2] designed for object detection. Here we note the final text embeddings of the categories in the target dataset (seen categories) as $\{T_k\}_{k=1}^{K}$, where K is the number of categories.

3.2. Grid-level alignment

Generating grid-level image embedding. Taking ResNet50 encoder as an example, in FCOS [27], the output feature maps C_3, C_4, C_5 of ResNet50 are inputted into FPN, producing 5 multi-scale image feature maps $\{P_i\}_{i=3}^{7}$. In FPN, C_5 is fused with C_3, C_4 to produce P_3, P_4 as well as serving as the input of P_6, P_7. Therefore, we fuse z into C_5 to propagate the image embeddings that are suitable for the text alignment across image feature maps of different scales.

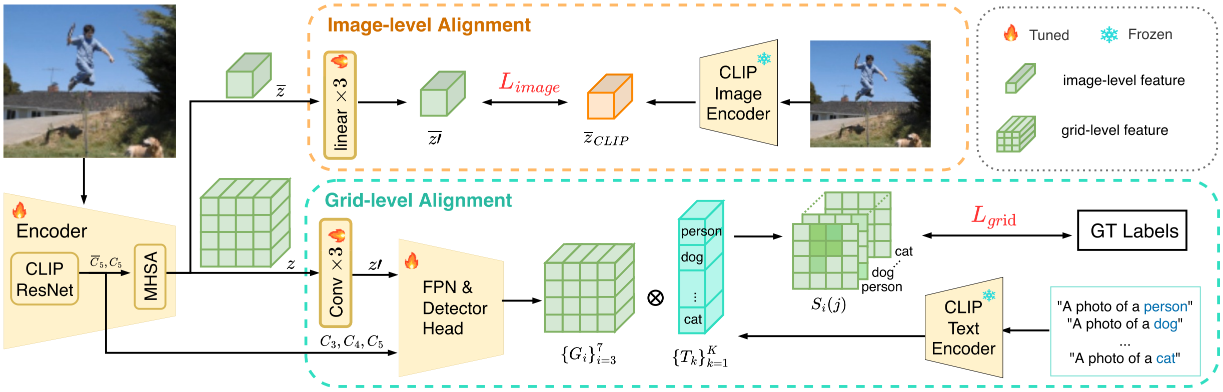


Fig. 2. The pipeline of our proposed GridCLIP. GridCLIP aligns with the CLIP representation in both image and grid levels. In image-level alignment, the image-level feature is aligned to the feature generated by a fixed CLIP image encoder. In grid-level alignment, the grid-level feature is aligned to the classification target generated from the ground truth labels and bounding boxes following the detector FCOS [27], we ignore the regression branch which remains the same as in FCOS for simplicity. Note that the grid-level alignment is performed in multiple scales, while only one scale is presented here for simplicity.

Specifically, we first apply three consecutive 3×3 convolutional layers with ReLU activation function to adapt the MHSA grid-level output feature z, reducing the number of channels from 1024 to 256 to generate z', and concatenate it with C_5. Then the concatenated feature $[z', C_5]$ replaces C_5 and is fed into the FPN with a little modification in the input channel number. In this way, the FPN is able to produce the multi-scale feature maps $\{P_i\}_{i=3}^{7}$ inheriting from the CLIP image embeddings that can be aligned to the text embedding as formulated below,

$$\{P_i\}_{i=3}^{7} = \mathrm{FPN}(C_3, C_4, [z', C_5]). \tag{2}$$

The FPN output features $\{P_i\}_{i=3}^{7}$ are then used to generate the final multi-scale grid-level features $\{G_i\}_{i=3}^{7}$ by going through the FCOS classification head. In the original FCOS, the classification head contains 5 convolutional layers and the last layer outputs the features with the channel number equal to the category number. However, in GridCLIP, we instead modify the output channel to be equal to the dimension of the CLIP text embeddings, to generate $\{G_i\}_{i=3}^{7}$. Then for each scale, the output feature map calculates the cosine similarities with each text embedding (each category) at the pixel level corresponding to the grids in the original image, to produce the multi-scale grid-level score with the Sigmoid activation function. For any grid (pixel in each feature map) j in the i-th scale grid-level feature $G_i(j)$, the matching score over all categories can be formulated as below,

$$S_i(j) = \left\{ \frac{G_i(j) \cdot T_k}{\|G_i(j)\|_2 \|T_k\|_2} \right\}_{k=1}^{K}. \tag{3}$$

Finally, the grid-level score $\{S_i\}_{i=3}^{7}$ is treated as the original classification output and aligned to the ground-truth target $\{Target_i\}_{i=3}^{7}$ using Focal Loss [26] as in original FCOS, noted as L_{grid}.

3.3. Image-level alignment

With grid-level alignment, the image grids of the seen categories are mapped to the CLIP alignment space, by aligning their embeddings to the corresponding text embeddings $\{T_k\}_{k=1}^{K}$. While for grids of unseen categories, there are no corresponding text embeddings to align to, which can only learn their embeddings by minimizing their similarity to any of $\{T_k\}_{k=1}^{K}$ during training. However, since the embeddings of different unseen categories can have different similarities to each seen category, simply minimizing the similarities between the seen and unseen categories is not consistent with the CLIP representation space which presents a generalizable knowledge representation. Therefore, ignoring the alignment of unseen categories may limit the ability to encode a wide range of unseen visual concepts, which harms the generalization ability of the model.

In practice, inspired by ViLD [2], we align the image-level embedding $\overline{z}$ to the embedding $\overline{z}_{\mathrm{CLIP}}$ produced by a fixed CLIP image encoder, so that the regions of unseen categories in an image can also be projected to the CLIP alignment space. Unlike ViLD, which requires multiple passes through the image encoder to align region embeddings from a separate RPN (making it a two-stage detector), our approach directly aligns the whole image embedding $\overline{z}'$ in a single pass, significantly reducing computational costs. Specifically, similar to grid-level alignment, we generate the image-level embedding $\overline{z'}$ by passing $\overline{z}$ through three consecutive linear layers with the ReLU activation function. Then we minimize the L_1 distance between $\overline{z'}$ and $\overline{z}_{\mathrm{CLIP}}$, which serves as the image-level alignment loss L_{image}. As for the fixed CLIP image encoder, we evaluate different published versions of pretrained models and choose the `ViT-B/32` version for alignment. Note that image-level alignment is only performed during the training phase.

Finally, the total loss for end-to-end training is:

$$L = w_{\mathrm{grid}} L_{\mathrm{grid}} + w_{\mathrm{image}} L_{\mathrm{image}} + L_{\mathrm{R}} + L_{\mathrm{C}}, \tag{4}$$

which includes the loss of two alignments and the original loss in the one-stage detector FCOS: regression loss L_{R} for bounding boxes and centerness loss L_{C} indicating the distance of a pixel to the center of the bounding box.

4. Experiments

4.1. Implementational details

GridCLIP uses the one-stage detector FCOS [27] as the baseline detector, which can be replaced by other one-stage detectors like RetinaNet [26]. The backbone uses the `RN50` pretrained CLIP image encoder, which has two more convolutional layers than the original ResNet50 [31] in the stem module and a Multi-Head Self-Attention (MHSA) layer applied to the output of the 5th stage. The adapting layers performed on the output features of MHSA use the embedding dimension of 256. For image-level alignment, GridCLIP uses the `ViT-B/32` version of CLIP. The weights of the two alignment losses are: $w_{\mathrm{grid}} = 1$, $w_{\mathrm{image}} = 10$. Our implementation is based on the MMDetection framework [34].

We conduct experiments on the detection benchmark LVIS v1.0 [35]. LVIS v1.0 is a long-tail detection dataset containing 1203 categories. The categories are divided into three parts by how many images they appear in: rare (1–10), common (11–100), and frequent (>100), respectively including 337, 461 and 405 categories, with corresponding Average Precision: AP_r, AP_c, and AP_f. Following ViLD [2], we use frequent and common categories as the seen categories for training. For open-vocabulary detection, rare categories are used as the unseen categories. We adopt multi-scale training similar to ViLD and random cropping augmentation. For the training process, GridCLIP is trained for a 2× (24 LVIS epochs) schedule with a batch size of 16. During the inference stage, the maximum number of detection objects per image is 300, and the threshold of the classification score is set to 0.05. The IOU threshold of NMS is 0.5.

4.2. Comparison with the state-of-the-art

We compare GridCLIP on the LVIS v1.0 validation set with other methods with comparable backbone, including ViLD [2], RegionCLIP [5], Detic [3], DetPro [6] and PromptDet [7] in Table 1. These methods use template-based prompts, except that DetPro and PromptDet use learnable prompts. Also, we do not compare to methods like GLIP [20] and DetCLIP [22] which use large-scale annotation data, since we focus on utilizing limited annotation data for the detection of broader categories.

Performance comparison. A totally fair comparison is not realistic, since external datasets or learnable prompts are widely used in most OVOD methods. Therefore, we find it relatively fair to compare GridCLIP with ViLD [2] which only utilizes the knowledge of CLIP and the detection dataset without learnable prompts. We observed that GridCLIP surpasses ViLD in overall AP by 0.8. As a one-stage detector, GridCLIP closes the gap to the two-stage detector ViLD in unseen categories to 1.3 AP_r, while the current SOTA one-stage detector HierKD is still 8.2 AP behind ViLD on the COCO validation dataset. Furthermore, GridCLIP outperforms ViLD by 1.5 AP_c and 0.9 AP_f. Besides `ViT-B/32`, we also use the `RN50×64` version (the largest model of CLIP under ResNet architecture) of CLIP for image-level alignment to explore the upper bound of the ResNet version. We observed that the `RN50×64` version has a worse generalization ability to unseen categories compared to the `ViT-B/32` one, with lower AP_r while higher AP_c and AP_f. We further observed the obvious gap between the seen and unseen categories in GridCLIP, and try to understand and explain the gap based on the analysis from another one-stage detector HierKD [12]. In ViLD, both the unseen and seen categories utilize the region-level alignment. In GridCLIP, whilst unseen categories use coarse-grained image-level alignment, the seen categories in GridCLIP use *both* fine-grained grid-level alignment and image-level alignment. This makes the gap between seen and unseen categories larger than that of ViLD. To verify this, we can replace

Table 1
Comparison with different object detectors on LVIS v1.0 [35] with open-set settings. Multi-scale training is used. "CLIP on cropped regions" directly applies CLIP to classify cropped region proposals. Except for GridCLIP, all detectors use an RPN pretrained on base categories to get region proposals. ‡ denotes using mask annotations. † denotes mask AP. ∇ denotes using learnable prompts instead of template prompts. ⋆ denotes that RegionCLIP use extra pretraining process of 600k iter on CC3M dataset with batch size of 96.

Model	Backbone	Pretrained CLIP	Epochs	External dataset	AP_r	AP_c	AP_f	AP
RegionCLIP [5]	CLIP R50-C4	RN50	12⋆	✓	17.1	27.4	34.0	28.2
Detic‡† [3]	R50-FPN	ViT-B/32	384	✓	17.8	26.3	31.6	26.8
PromptDet∇† [7]	R50-FPN	ViT-B/32	6+12	✓	19.0	18.5	25.8	21.4
CLIP on cropped regions‡ [2]	R50-FPN	ViT-B/32	0	✗	**19.5**	19.7	17.0	18.6
ViLD‡ [2]	R50-FPN	ViT-B/32	384	✗	<u>16.3</u>	21.2	31.6	24.4
GridCLIP-R50	CLIP R50-FPN	ViT-B/32	24	✗	15.0	<u>22.7</u>	<u>32.5</u>	<u>25.2</u>
GridCLIP-R50-RN	CLIP R50-FPN	RN50x64	24	✗	13.7	**23.3**	**32.6**	**25.3**

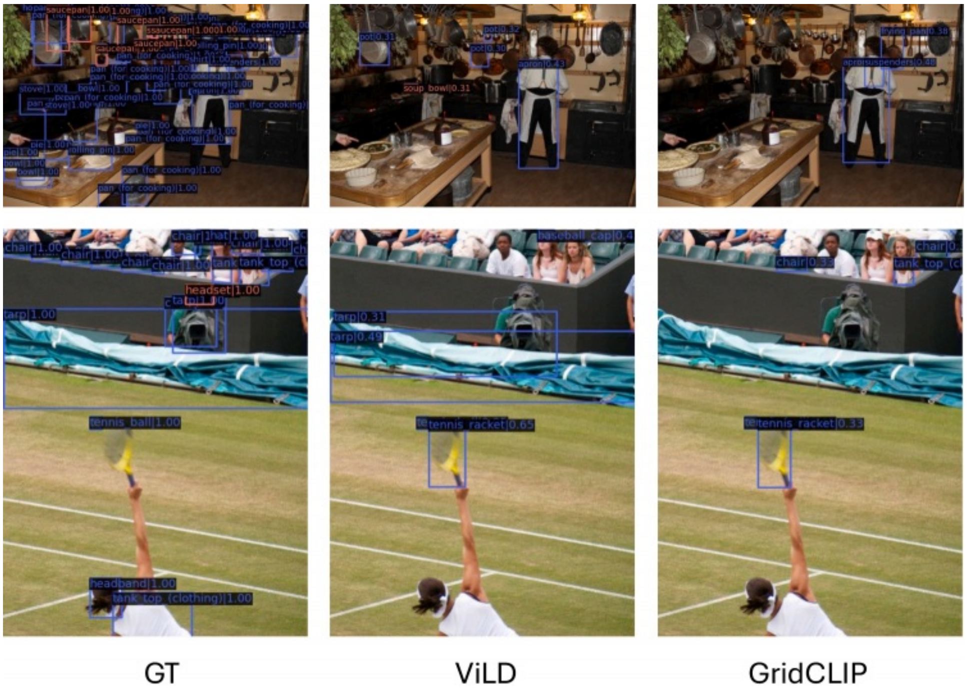


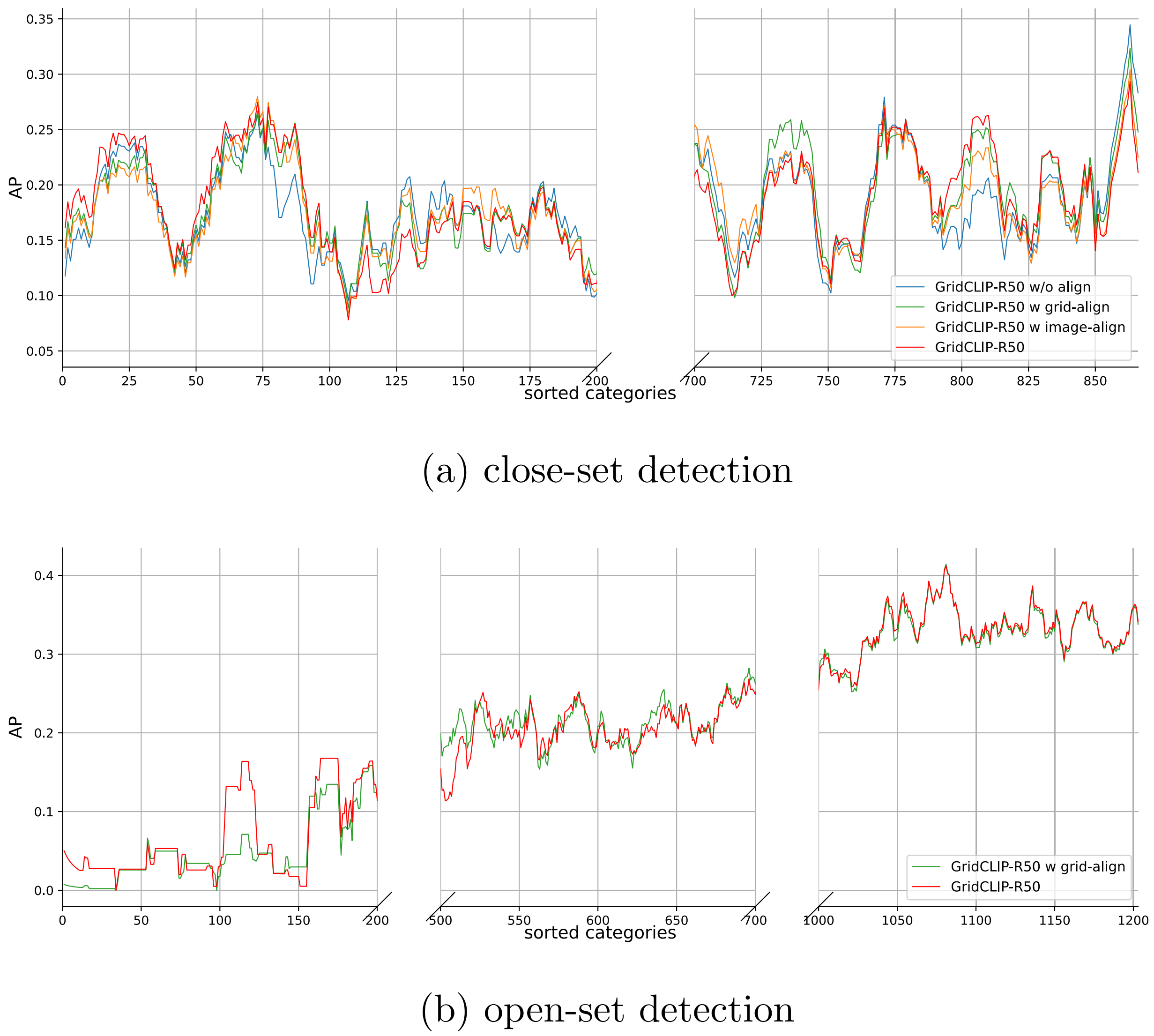
Fig. 3. Qualitative comparisons with ViLD [2], using a visualization threshold of 0.3.

image-level alignment with region-level alignment similar to ViLD to further improve AP_r, which however may require more training time as other two-stage detectors do. Furthermore, the qualitative comparison is shown in Fig. 3. ViLD and GridCLIP demonstrate comparable performance with complementary strengths, effectively detecting prominent objects. For smaller instances (e.g., pot, chair), both models show detection capability. However, GridCLIP occasionally misses large, salient objects such as the tarp. This limitation may stem from the lack of region-level feature extraction as in two-stage detectors, which aggregate contextual information and enhances the receptive field.

Computation efficiency. We compare GridCLIP with ViLD on training and test time, as well as the model size. ViLD is originally trained on TPUv3 with a batch size of 256. For fair comparison and due to resource limitation, we train both ViLD and GridCLIP with a batch size of 16 on 2 A100 GPUs. We train ViLD using the implementation of DetPro [6]. Moreover, ViLD takes 1 day on 8 V100 GPUs to pre-compute the CLIP image embedding of regions to accelerate training. GridCLIP does not require this. In Table 2, we show that with comparable model size, GridCLIP-R50 is approximately 43 times and 5 times faster than ViLD in training and test respectively. Such significant advantages remain when GridCLIP-R50-RN using `RN50×64` (3 times larger in both input and parameters) for image-level alignment, with 34 times faster in training time than that of ViLD. This validates clearly the computational efficiency of the one-stage GridCLIP.

Table 2
The training and test time on LVIS v1.0 and the model size comparisons of ViLD and GridCLIP. The resource usage for ViLD is measured based on the implementation in DetPro [6]. Note that we use the original AP reported in the ViLD paper due to the significant computational resources required for training. Our one-stage method performs comparable performance with two-stage ViLD with 43× less hours.

Model	Parameters for Inference (M)	Epoch	Training Cost / Epoch (Per-GPU-Hour)	Total Training Cost (Per-GPU-Hour)	FPS	AP_r	AP
ViLD [2]	60.5	384	7.98	3064	3.3	16.3	24.4
GridCLIP-R50	56.4	24	2.94	70	19.5	15.0	25.2
GridCLIP-R50-RN	56.4	24	3.66	88	19.5	13.7	25.3



(a) close-set detection

(b) open-set detection

Fig. 4. The AP on LVIS v1.0 over categories sorted by frequency in ascending order. (a) uses on the close-set setting, only containing 461 “common” categories and 405 “frequent” categories. (b) uses the open-set setting, containing containing 337 “rare” categories, 461 “common” categories and 405 “frequent” categories. The value is smoothed using moving average with window [-10,10].

Transfer to other datasets. To further explore the generalizability of GridCLIP, we follow ViLD [2] and evaluate the LVIS-trained GridCLIP on both the PASCAL VOC 2007 test set [37] and the COCO validation set [38] by directly replacing the categories without any finetuning. Note that there are overlaps of both category and image between LVIS and COCO (as well as PASCAL VOC). The IOU threshold of NMS is 0.6. On PASCAL VOC, we observed that the gap between GridCLIP-R50 and ViLD is 1.2 to 1.3 AP, and GridCLIP-R50-RN is comparable with ViLD on PASCAL VOC with no more than 1 AP difference. Although the gap on COCO is still obvious with 2.2 AP falling behind but is comparable to that of DetPro which uses learn prompts based on ViLD. Therefore, in the generalization ability, GridCLIP performs quite close to its two-stage counterparts (Table 3).

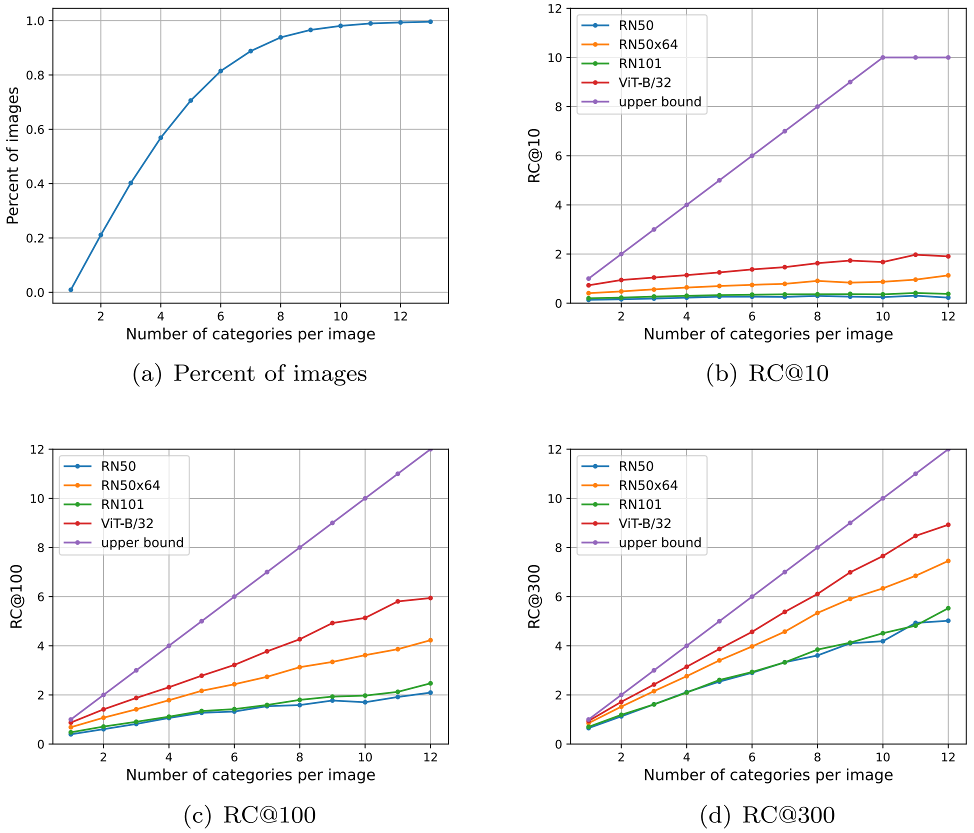
Table 3
Generalization ability of LVIS-trained detectors to PASCAL VOC 2007 test set and COCO validation set. Note that both ViLD and DetPro use extra mask annotations for supervison during training, while GridCLIP only use bounding box supervison.

Model	PASCAL VOC		COCO		
	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
ViLD	72.2	56.7	36.6	55.6	39.8
DetPro	74.6	57.9	34.9	53.8	37.4
GridCLIP-R50	72.1	55.2	34.3	52.4	36.0
GridCLIP-R50-RN	71.4	54.9	34.6	52.9	36.5

4.3. Ablation studies

We verify the effectiveness of grid-level and image-level alignment for minority categories for both closed-vocabulary detection (Table 4) and open-vocabulary detection (Table 5). We follow the same settings in Section 4.2 except that multi-scaling training and random cropping augmentation are not used here. Among the experiments, “w/o align” denotes using the original design of FCOS only with different backbones that feed different image features to the FPN. “w grid-align” and “w image-align” denote only using grid-level or image-level alignment, respectively.

Closed-vocabulary detection. For closed-vocabulary detection, we train and test only on the common and frequent categories, containing 866 categories. We first evaluate other visual pretrained models as the backbone to compare to the vision-language pretrained model



(a) Percent of images

(b) RC@10

(c) RC@100

(d) RC@300

Fig. 5. (a) shows the percent of images containing different numbers of categories in an image on the LVIS v1.0 validation set. (b), (c) and (d) are the recall of top k (k = 10, 100, 300) predictions of different CLIP pretrained versions, using the image-level representation from the corresponding CLIP image encoder. These figures evaluate the capacity of the original CLIP image representation in representing multiple categories in an image, and further verify that our one-stage detector GridCLIP built upon the CLIP image-level representation can benefit image-level alignment.

CLIP, including the ImageNet [30] pretrained ResNet50 on the classification task and self-supervised pretrained ResNet50 using visual-only SSL method SwAV [36] pretrained on large-scale unlabeled images. By comparing the methods using different pretrained ResNet50 without any alignment (the top section of Table 4), we notice that using the CLIP pretrained ResNet50 can bring notable improvements compared to the ImageNet and SwAV pretrained ones, with the similar architecture, which indicates the superiority of the vision-language pretrained model CLIP than other visual pretrained models in generalizing better image embeddings for detection.

On the bottom section of Table 4, we introduce the MHSA layer whose output is aligned to the CLIP text encoder in the original training of CLIP. Therefore, the MHSA layer provides both grid-level and image-level features for CLIP-based alignment. We first observed that introducing the MHSA layer without any alignment drops the overall performance by 0.7 AP, while using both alignments can improve the performance in the infrequent common categories by 1.3 AP and preserve the performance in frequent categories. Among the experiments that use the MHSA layer, we find that applying grid-level alignment can significantly improve the common categories by 2.3 AP, while using image-level alignment has limited impact on the metrics AP_c and AP_f. Using both alignments can bring 1.8 AP_c and 0.3 AP_f improvements which are lower than the one using only grid-level alignment. To further explore the reasons behind that, we first find that the metrics of AP_c and AP_f are too coarse-grained for distinguishing categories with different frequencies, which can suffer from the bias in the way of splitting the dataset. So we present the performance in a more fine-grained way by the plot of AP over categories with different sample numbers (Fig. 4). As shown in subfigure (a), in the 200 most infrequent categories, the improvement of applying one alignment is not stable, where

Table 4
Comparison of different backbones and alignment methods on LVIS v1.0 [35] with close-set settings. The top section compares supervised (ImageNet [30]) and unsupervised (SwAV [36]) pretrained visual models with CLIP as unsupervised visual-language pretrained model. The bottom section compares different alignment methods based on the ResNet50 version of pretrained CLIP image encoder. “GridCLIP-R50∗” uses the CLIP ResNet50 without the MHSA layer.

Model	Backbone	Grid-level Alignment	Image-level Alignment	AP_c	AP_f	AP
ImageNet-R50 w/o align	ImageNet R50-FPN	-	-	14.6	26.2	16.6
SwAV-R50 w/o align	SwAV R50-FPN	-	-	19.5	29.2	20.0
GridCLIP-R50∗ w/o align	CLIP R50-FPN woMHSA	-	-	20.2	**30.3**	20.7
GridCLIP-R50 w/o align	CLIP R50-FPN	✗	✗	19.4	29.7	20.1
GridCLIP-R50 w grid-align	CLIP R50-FPN	✓	✗	**21.7**	30.0	**21.2**
GridCLIP-R50 w image-align	CLIP R50-FPN	✗	✓	19.4	30.1	20.2
GridCLIP-R50	CLIP R50-FPN	✓	✓	21.2	30.0	21.0

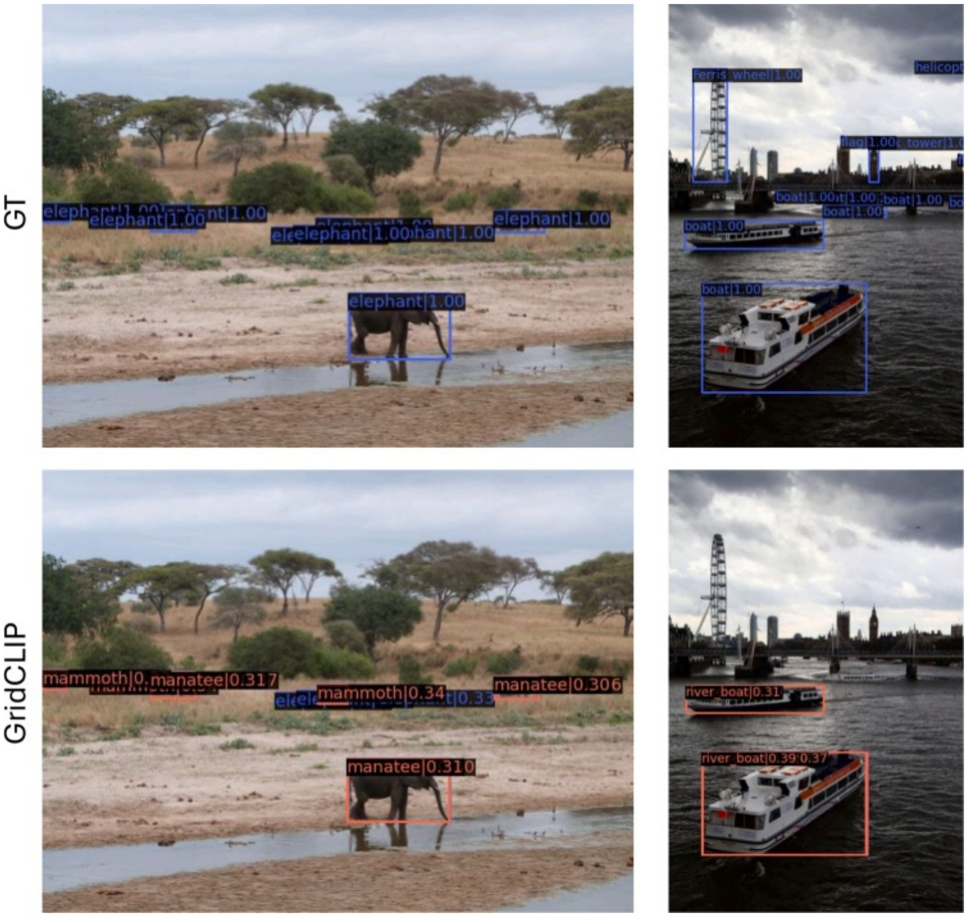


Fig. 6. Failure cases.

Table 5
The effectiveness of image-level alignment on LVIS v1.0 [35] under open-set settings.

Model	AP_r	AP_c	AP_f	AP
GridCLIP-R50 w grid-align	10.1	**21.0**	29.6	22.5
GridCLIP-R50	**12.7**	20.6	**29.7**	**22.8**

the AP can be notably higher than “GridCLIP-R50 w/o align” in some categories while obviously worse in other categories. When applying both alignments, “GridCLIP-R50” achieves top performance primarily in the 100 most infrequent categories, though its superiority is less pronounced in frequent categories. Therefore, we can conclude that using both alignments benefits the minority categories in closed-vocabulary detection.

Open-vocabulary detection. Since only models with grid-level alignment can be extended for open-vocabulary detection by extending the categories embedding list with unseen categories and using the list match with the grid-level image embeddings, we compare two models from Table 4. As shown in Table 5, image-level alignment improves unseen categories by 2.6 AP_r, while preserving the performance on seen categories. For analysis of category-wise accuracy across sample frequency spectrum, in Fig. 4 (b), we observed that the performance rises over categories as their training sample number increases, which verifies that minority categories suffer from long-tail distribution. In the 200 most infrequent categories, which are unseen (also rare) categories, “GridCLIP-R50” outperforms “GridCLIP-R50 w grid-align” significantly, which indicates the effectiveness of image-level alignment on unseen categories. Therefore, it is verified that the alignment of image-level representations also helps learn generalizable grid-level representations of minority categories.

In summary, grid-level alignment enhances detection of minority categories in closed-vocabulary tasks and enables open-vocabulary detection. Image-level alignment benefits unseen categories in open-vocabulary detection. Combining both alignments allows one-stage detectors to identify unseen categories and improve performance on minority categories in long-tail datasets.

4.4. Further analysis: CLIP image-level representation for object detection

We analyse how accurately the multiple categories in an image can be represented by the original CLIP image encoder. This substantially affects the performance of a one-stage detector built upon the CLIP image-level representation to detect multiple categories at the same time, which indicates how much image-level alignment can benefit GridCLIP. We evaluate several pretrained versions of CLIP on LVIS v1.0 validation set, including the refined ResNet [31] (`RN50, RN101, RN50×64`) and those with the Transformer architecture [32] (`ViT-B/32`). Specifically, we calculate the recall of categories by using the original CLIP image-level representation to match the text representation of each category (Fig. 5).

For RC@10, all models perform poorly with no more than 2 recalls. While for RC@100, we find that `ViT-B/32` can recall 50 % of the categories and `RN50×64` can recall more than 30 % of the categories. In comparison, the other ResNet-based models perform poorly that reach less than 20 % recall rate. Furthermore, for RC@300, nearly 75 % of the categories in an image are captured by `ViT-B/32`, and `RN50×64` reaches about 60 % recall rate. Given that the maximum detection number for LVIS v1.0 in OVOD is usually set to 300 (objects), `ViT-B/32` can at most help detect 75 % of the objects if all the objects have different categories and 50 % if every 3 of the objects share the same category. This provides substantial knowledge of categories to help the detector build the representation for multiple object detection. Therefore, the CLIP image encoder is able to capture multiple categories in an image at the same time with relatively high accuracy and provide substantial knowledge of categories for the detector during image-level alignment.

4.5. Failure cases

We illustrate representative failure cases of GridCLIP in Fig. 6. The model occasionally confuses visually similar categories, such as misclassifying the seen category “elephant” as the unseen category “mammoth.” This phenomenon can be attributed to: (1) the extreme scarcity of mammoth instances in the LVIS v1.0 dataset (with only up to 10 annotated images), and (2) the pre-trained CLIP encoder’s tendency to map related concepts to nearby embeddings, resulting in an ambiguous decision boundary between these semantically and visually related classes. Such challenges are common for long-tailed data distributions, where insufficient representation of minority categories impedes the model’s ability to distinguish fine-grained differences. Another observed failure involves the misclassification of the seen category “boat” as the unseen category “river_boat” (as shown in the second column). This error is primarily attributable to ambiguous or imprecise category annotations within LVIS v1.0, where “river_boat” arguably provides a more specific and accurate label than the broader “boat” category. Such cases highlight the impact of annotation granularity and label quality on open-vocabulary detection performance.

5. Conclusion

We present GridCLIP, a one-stage detector that leverages CLIP’s representation space to enhance minority category detection. It optimizes CLIP’s pretrained knowledge for fine-grained localization through two mechanisms: grid-level alignment that maps localized features to seen categories, and image-level alignment that performs holistic knowledge distillation for both seen and unseen categories. GridCLIP demonstrates improved performance on long-tail distributions, achieving comparable results to state-of-the-art methods on LVIS v1.0 with faster training and inference, serving as a strong baseline for future downstream applications, like multi-object tracking [39,40].

6. Limitations and future work

While GridCLIP achieves promising results, it still exhibits an accuracy gap on unseen categories (1.3 AP_r behind ViLD). This gap largely stems from the reliance on coarser image-level alignment for unseen categories, in contrast to the fine-grained region-level alignment employed by two-stage detectors. Addressing this imbalance is an important direction for future work. One possible approach is to incorporate finer-grained alignment for unseen categories. Ideally, this would mirror the proposal-level alignment of two-stage detectors; while this offers rich supervision, it is computationally expensive. To retain efficiency, a promising alternative is to perform sparse region alignment without relying on dense proposals for example, by selecting a small number of discriminative regions and aligning them through the CLIP image encoder. We believe that designing lightweight yet effective alignment strategies can further enhance unseen-category performance while preserving the simplicity and efficiency of the one-stage framework.

CRediT authorship contribution statement

Jiayi Lin: Writing – review & editing, Writing – original draft, Visualization, Validation, Project administration, Methodology, Formal analysis, Data curation, Conceptualization; **Shitong Sun:** Writing – review & editing; **Shaogang Gong:** Writing – review & editing, Supervision, Conceptualization.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A.I. Saichev, Y. Malevergne, D. Sornette, Theory of Zipf’s Law and Beyond, 632, Springer Science & Business Media, 2009.

[2] X. Gu, T.-Y. Lin, W. Kuo, Y. Cui, Open-vocabulary object detection via vision and language knowledge distillation, in: International Conference on Learning Representations, 2022. https://www.openreview.net/forum?id=lL3lnMbR4WU.

[3] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, I. Misra, Detecting twenty-thousand classes using image-level supervision, in: European Conference on Computer Vision, Springer, 2022.

[4] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

[5] Y. Zhong, J. Yang, P. Zhang, C. Li, N. Codella, L.H. Li, L. Zhou, X. Dai, L. Yuan, Y. Li, J. 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, IEEE, 2022, pp. 16772–16782. https://doi.org/10.1109/CVPR52688.2022.01629

[6] Y. Du, F. Wei, Z. Zhang, M. Shi, Y. Gao, G. Li, Learning to prompt for open-vocabulary object detection with vision-language model, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 14084–14093.

[7] C. Feng, Y. Zhong, Z. Jie, X. Chu, H. Ren, X. Wei, W. Xie, L. Ma, Promptdet: towards Open-vocabulary detection using uncurated images, in: European Conference on Computer Vision, Springer, 2022, pp. 701–717.

[8] W. Kuo, Y. Cui, X. Gu, A.J. Piergiovanni, A. Angelova, Open-vocabulary object detection upon frozen vision and language models, in: The Eleventh International Conference on Learning Representations, 2023. https://www.openreview.net/forum?id=MIMwy4kh9lf.

[9] Z. Piao, J. Wang, L. Tang, B. Zhao, W. Wang, Accloc: anchor-free and two-stage detector for accurate object localization, Pattern Recognit. 126 (2022) 108523.

[10] J. Xie, S. Zheng, Zero-shot object detection through vision-language embedding alignment, in: 2022 IEEE international conference on data mining workshops (ICDMW), IEEE, 2022, pp. 1–15.

[11] Y. Rao, W. Zhao, G. Chen, Y. Tang, Z. Zhu, G. Huang, J. Zhou, J. Lu, DenseCLIP: language-guided dense prediction with context-aware prompting, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18–24, 2022, IEEE, 2022, pp. 18061–18070. https://doi.org/10.1109/CVPR52688.2022.01755

[12] Z. Ma, G. Luo, J. Gao, L. Li, Y. Chen, S. Wang, C. Zhang, W. Hu, Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 14074–14083.

[13] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, Y. Shan, Yolo-world: real-time open-vocabulary object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 16901–16911.

[14] H. Cheng, H. Ye, X. Zhou, X. Liu, F. Chen, M. Wang, Vision-language pre-training via modal interaction, Pattern Recognit. 156 (2024) 110809.

[15] C. Zhou, C.C. Loy, B. Dai, Extract free dense labels from clip, in: European Conference on Computer Vision, Springer, 2022, pp. 696–712.

[16] S. Sun, C. Si, G. Wu, S. Gong, Federated zero-shot learning with mid-level semantic knowledge transfer, Pattern Recognit. 156 (2024) 110824.

[17] F. Zheng, J. Cao, W. Yu, Z. Chen, N. Xiao, Y. Lu, Exploring low-resource medical image classification with weakly supervised prompt learning, Pattern Recognit. 149 (2024) 110250.

[18] Q. Yan, Y. Yang, Y. Dai, X. Zhang, K. Wiltos, M. Woźniak, W. Dong, Y. Zhang, CLIP-guided continual novel class discovery, Knowl. Based Syst. 310 (2025) 112920.

[19] W. Cai, J. Huang, S. Gong, H. Jin, Y. Liu, MLLM as video narrator: mitigating modality imbalance in video moment retrieval, Pattern Recognit. 166 (2025) 111670.

[20] L.H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang, et al., Grounded language-image pre-training, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10965–10975.

[21] A. Zareian, K.D. Rosa, D.H. Hu, S.-F. Chang, Open-vocabulary object detection using captions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14393–14402.

[22] H. Zhang, P. Zhang, X. Hu, Y.-C. Chen, L. Li, X. Dai, L. Wang, L. Yuan, J.-N. Hwang, J. Gao, Glipv2: unifying localization and vision-language understanding, Adv. Neural Inf. Process. Syst. 35 (2022) 36067–36080.

[23] L. Yao, J. Han, Y. Wen, X. Liang, D. Xu, W. Zhang, Z. Li, C. Xu, H. Xu, DetCLIP: dictionary-enriched visual-concept paralleled pre-training for open-world detection, in: A.H. Oh, A. Agarwal, D. Belgrave, K. Cho (Eds.), Advances in Neural

Information Processing Systems, 2022. https://www.openreview.net/forum?id=4rTN0MmOvi7.

[24] K. Zhou, J. Yang, C.C. Loy, Z. Liu, Learning to prompt for vision-language models, Int. J. Comput. Vis. 130 (9) (2022) 2337–2348. https://doi.org/10.1007/s11263-022-01653-1

[25] S. Fu, Q. Yang, Q. Mo, J. Yan, X. Wei, J. Meng, X. Xie, W.-S. Zheng, LLMDet: learning strong open-vocabulary object detectors under the supervision of large language models, arXiv preprint arXiv:2501.18954. (2025).

[26] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.

[27] Z. Tian, C. Shen, H. Chen, T. He, Fcos: fully convolutional one-stage object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9627–9636.

[28] M. Gao, C. Xing, J.C. Niebles, J. Li, R. Xu, W. Liu, C. Xiong, Open vocabulary object detection with pseudo bounding-box labels, in: European Conference on Computer Vision, Springer, 2022, pp. 266–282.

[29] P. Sharma, N. Ding, S. Goodman, R. Soricut, Conceptual captions: a cleaned, hypernymed, image alt-text dataset for automatic image captioning, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 2556–2565.

[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255.

[31] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[32] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. 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. https://www.openreview.net/forum?id=YicbFdNTTy.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[34] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, et al., MMDetection: open mmlab detection toolbox and benchmark, arXiv preprint arXiv:1906.07155. (2019).

[35] A. Gupta, P. Dollar, R. Girshick, LVIS: a dataset for large vocabulary instance segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5356–5364.

[36] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, A. Joulin, Unsupervised learning of visual features by contrasting cluster assignments, Adv. Neural Inf. Process. Syst. 33 (2020) 9912–9924.

[37] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338.

[38] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: common objects in context, in: European Conference on Computer Vision, Springer, 2014, pp. 740–755.

[39] Y. Hu, A. Niu, J. Sun, Y. Zhu, Q. Yan, W. Dong, M. Woźniak, Y. Zhang, Dynamic center point learning for multiple object tracking under severe occlusions, Knowl. Based Syst. 300 (2024) 112130.

[40] Y. Hu, J. Sun, H. Jin, A. Niu, Q. Yan, Y. Zhu, Y. Zhang, Robust multi-object tracking using vision sensor with fine-grained cues in occluded and dynamic scenes, IEEE Sens. J. (2025).

Author biographies

Jiayi Lin is a Ph.D. Candidate in Computer Science at Queen Mary University of London, supervised by Prof. Shaogang Gong. Her research interest includes image recognition, Computer Vision (Low-level Vision) and Deep Learning. She received her master’s degree at School of Artificial Intelligence, University of Chinese Academy of Sciences in 2021, supervised by Prof. Liang Wang. And in 2018, she received her bachelor’s degree from Sun Yat-sen University.

Shitong Sun received her Ph.D. in Computer Vision at Queen Mary University of London under the supervision of Prof. Shaogang Gong. She previously earned a BEng degree from KU Leuven in 2017 and MSc degree in KU Leuven in 2018. Her current research interests include multimodal and federated learning.

Shaogang Gong is professor of Visual Computation at Queen Mary University of London; elected a Fellow of the Royal Academy of Engineering (FREng), a Fellow of ELLIS, a Fellow of AAIA, a Fellow of the Institution of Electrical Engineers, a Fellow of the British Computer Society, a member of the UK Computing Research Committee, a Turing Fellow of the Alan Turing Institute, and served on the Steering Panel of the UK Government Chief Scientific Advisor’s Science Review. He received the D.Phil. degree in computer vision from the Keble College, Oxford University, in 1989. He has published more than 400 research articles and seven books on topics, including Person Re-Identification, Visual Analysis of Behavior, Video Analytics for Business Intelligence, Dynamic Vision: From Images to Face Recognition, and Analysis and Modeling of Faces and Gestures. He is the inventor of 42 international patents. His research interests include computer vision, machine learning, and video analysis. He served on the Steering Panel of the U.K. Government Chief Scientific Advisor’s Science Review. He won the Institution of Engineering and Technology 2020 Achievement Medal for Vision Engineering for outstanding achievement and superior performance in contributing to public safety.