Grounded Language-Image Pre-training

Liunian Harold Li[*1†], **Pengchuan Zhang**[*2♠], **Haotian Zhang**[*3†], **Jianwei Yang**[2], **Chunyuan Li**[2], **Yiwu Zhong**[4†], **Lijuan Wang**[5], **Lu Yuan**[5], **Lei Zhang**[6], **Jenq-Neng Hwang**[3], **Kai-Wei Chang**[1], **Jianfeng Gao**[2]

[1]UCLA, [2]Microsoft Research, [3]University of Washington,

[4]University of Wisconsin-Madison, [5]Microsoft Cloud and AI, [6]International Digital Economy Academy

Abstract

This paper presents a grounded language-image pre-training (GLIP) model for learning object-level, language-aware*, and* semantic-rich *visual representations. GLIP unifies object detection and phrase grounding for pre-training. The unification brings two benefits: 1) it allows GLIP to learn from both detection and grounding data to improve both tasks and bootstrap a good grounding model; 2) GLIP can leverage massive image-text pairs by generating grounding boxes in a self-training fashion, making the learned representations semantic-rich. In our experiments, we pre-train GLIP on 27M grounding data, including 3M human-annotated and 24M web-crawled image-text pairs. The learned representations demonstrate strong zero-shot and few-shot transferability to various object-level recognition tasks. 1) When* directly evaluated *on COCO and LVIS (without seeing any images in COCO during pre-training), GLIP achieves 49.8 AP and 26.9 AP, respectively, surpassing many supervised baselines.*[1] *2) After* fine-tuned *on COCO, GLIP achieves 60.8 AP on val and 61.5 AP on test-dev, surpassing prior SoTA. 3) When* transferred *to 13 downstream object detection tasks, a 1-shot GLIP rivals with a fully-supervised Dynamic Head. Code will be released at* https://github.com/microsoft/GLIP.

1. Introduction

Visual recognition models are typically trained to predict a fixed set of pre-determined object categories, which limits their usability in real-world applications since additional labeled data are needed to generalize to new visual concepts and domains. CLIP [40] shows that *image-level* visual representations can be learned effectively on large amounts of raw image-text pairs. Because the paired texts contain a boarder set of visual concepts than any pre-defined concept pool, the pre-trained CLIP model is so semantically rich that it can be easily transferred to downstream image classification and text-image retrieval tasks in zero-shot settings. However, to gain fine-grained understanding of images, as required by many tasks, such as object detection [31, 44], segmentation [6, 35], human pose estimation [49, 56], scene understanding [14, 25, 57], action recognition [17], vision-language understanding [7, 27–30, 36, 48, 50, 63, 65], *object-level* visual representations are highly desired.

In this paper, we show that *phrase grounding*, which is a task of identifying the fine-grained correspondence between phrases in a sentence and objects (or regions) in an image, is an effective and scalable pre-training task to learn an object-level, language-aware, and semantic-rich visual representation, and propose Grounded Language-Image Pre-training (GLIP). Our approach unifies the phrase grounding and object detection tasks in that object detection can be cast as context-free phrase grounding while phrase grounding can be viewed as a contextualized object detection task. We highlight our key contributions as follows.

Unifying detection and grounding by reformulating object detection as phrase grounding. The reformulation changes the input of a detection model: it takes as input not only an image but also a text prompt that describes *all* the candidate categories in the detection task[2]. For example, the text prompt for COCO object detection [32] is a text string that consists of 80 phrases, i.e., the 80 COCO object class names, joined by ". ", as shown in Figure 1 (Left). Any object detection model can be converted to a grounding model by replacing the object classification logits in its box classifier with the word-region alignment scores, i.e., dot product of the region (or box) visual features and the token (or phrase) language features, as shown in Figure 1 (Right). The language features are computed using a language model, which gives the new detection (or grounding) model a dual-encoder structure. Different from CLIP that fuses vision and language only at the last dot product layer [40], we show that deep cross-modality fusion applied

[*]The three authors contributed equally. [♠] Corresponding author.

[†]Work done when interning at Microsoft Research.

[1]Supervised baselines on COCO object detection: Faster-RCNN w/ ResNet50 (40.2) or ResNet101 (42.0), and DyHead w/ Swin-Tiny (49.7).

[2]Different from typical phrase grounding tasks, phrases in the text prompt for an object detection task may not be present in the image.

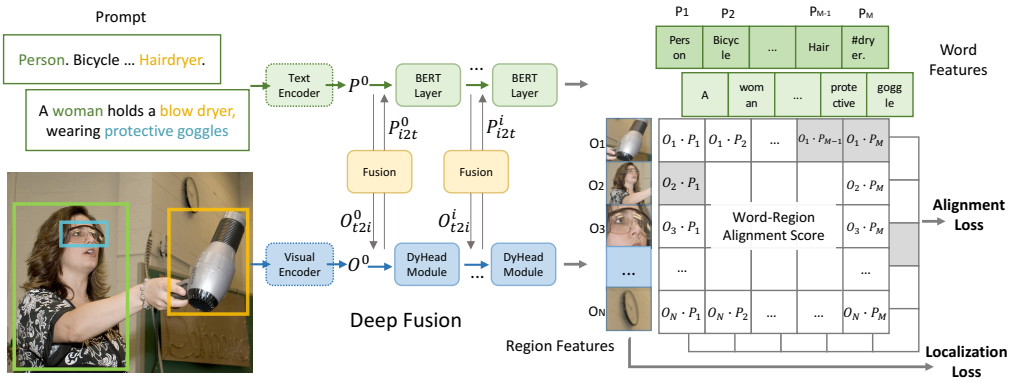

Figure 1. A unified framework for detection and grounding. Unlike a classical object detection model which predicts a categorical class for each detected object, we reformulate detection as a grounding task by aligning each region/box to phrases in a text prompt. GLIP jointly trains an image encoder and a language encoder to predict the correct pairings of regions and words. We further add the cross-modality deep fusion to early fuse information from two modalities and to learn a language-aware visual representation.

Figure 2. Grounding predictions from GLIP. GLIP can locate rare entities, phrases with attributes, and even abstract words.

by GLIP, as shown in Figure 1 (Middle), is crucial to learn high-quality language-aware visual representations and to achieve superior transfer learning performance. The unification of detection and grounding also allows us to pre-train using both types of data and benefits both tasks. On the detection side, the pool of visual concepts is significantly enriched thanks to the grounding data. On the grounding side, detection data introduce more bounding box annotations and help train a new SoTA phrase grounding model.

Scaling up visual concepts with massive image-text data. Given a good grounding model (teacher), we can augment GLIP pre-training data by automatically generating grounding boxes for massive image-text-paired data, in which noun phrases are detected by an NLP parser [2]. Thus, we can pre-train our (student) GLIP-Large model (GLIP-L) on 27M grounding data, including 3M human-annotated fine-grained data and 24M web-crawled image-text pairs. For the 24M image-text pairs, there are 78.1M high-confidence (> 0.5) phrase-box pseudo annotations, with 58.4M unique noun phrases. We showcase two real examples of the generated boxes in Figure 2. The teacher model can accurately localize some arguably hard concepts, such as *syringes*, *vaccine*, *beautiful caribbean sea turquoise*, and even abstract words (*the view*). Training on such semantic-rich data delivers a semantic-rich student model. In contrast, prior work on scaling detection data simply cannot predict concepts out of the teacher models' pre-defined vocabulary [68]. In this study, we show that this simple strategy of scaling up grounding data is empirically effective, bringing large improvements to LVIS and 13 downstream detection tasks, especially on rare categories (Sections 4.2 and 5). When the pre-trained GLIP-L model is fine-tuned on COCO, it achieves 60.8 AP on COCO 2017val and 61.5 on test-dev, surpassing the current public SoTA models [9, 58] that scale up object detection data in various approaches.

Transfer learning with GLIP: one model for all. The grounding reformulation and semantic-rich pre-training facilitate domain transfer. GLIP can be transferred to various tasks with few or even no additional human annotations. When the GLIP-L model is directly evaluated on the COCO and LVIS datasets (without seeing any images in COCO during pre-training), it achieves 49.8 and 26.9 AP on COCO val2017 and LVIS val, respectively, surpassing many supervised baselines. When evaluated on 13 existing object detection datasets, spanning scenarios including fine-grained species detection, drone-view detection, and ego-centric detection, the setting which we term "Object Detection in the Wild" (ODinW) (Section 5.1), GLIP exhibits excellent data efficiency. For example, a zero-shot GLIP-L outperforms a 10-shot supervised baseline (Dynamic Head) pre-trained on Objects365 while a 1-shot GLIP-L rivals with a fully supervised Dynamic Head. Moreover, when task-specific annotations are available, instead of tuning the whole model, one could tune only the task-specific prompt embedding, while keeping the model parameters unchanged. Under such a prompt tuning setting (Section 5.2), one GLIP model can simultaneously perform well on all downstream tasks , reducing the fine-tuning and deployment cost.

2. Related Work

Standard object detection systems are trained to localize a fixed set of object classes predefined in crowd-labeled datasets, such as COCO [32], OpenImages (OI) [25], Objects365 [45], and Visual Genome (VG) [23], which con-

tains no more than 2,000 object classes. Such human-annotated data are costly to scale up. GLIP presents an affordable solution by reformulating object detection as a phrase grounding (word-to-region matching) problem, and thus enables the use of grounding and massive image-text-paired data. Though our current implementation is built upon Dynamic Head (DyHead) [9], our unified formulation can be generalized to any object detection systems [4, 5, 8, 9, 9, 31, 43, 44, 67].

Recently, there is a trend to develop vision-and-language approaches to visual recognition problems, where vision models are trained with free-form language supervision. For example, CLIP [40] and ALIGN [18] perform cross-modal contrastive learning on hundreds or thousands of millions of image-text pairs and can directly perform open-vocabulary image classification. By distilling the knowledge from the CLIP/ALIGN model into a two-stage detector, ViLD [12] is proposed to advance zero-shot object detection. Alternatively, MDETR [19] trains an end-to-end model on existing multi-modal datasets which have explicit alignment between phrases in text and objects in image. Our GLIP inherits the semantic-rich and language-aware property of this line of research, achieves SoTA object detection performance and significantly improves the transferability to downstream detection tasks.

This paper focuses on domain transfer for object detection. The goal is to build one pre-trained model that seamlessly transfers to various tasks and domains, in a zero-shot or few-shot manner. Our setting differs from zero-shot detection [1, 12, 41, 42, 61, 66], where some categories are defined as unseen/rare and not present in the training set. We expect GLIP to perform well on rare categories (Section 4.2) but we do not explicitly exclude any categories from our training set, because grounding data are so semantically rich that we expect them to cover many rare categories. This resembles the setting in open-vocabulary object detection [61], which expects raw image-text data to cover many rare categories. Beyond performance on rare categories, we also consider the transfer cost in real-world scenarios, i.e., how to achieve the best performance with the least amount of data, training budget, and deployment cost (Section 5).

3. Grounded Language Image Pre-training

Conceptually, object detection and phrase grounding bear a great similarity. They both seek to localize objects and align them to semantic concepts. This synergy motivates us to cast the classical object detection task into a grounding problem and propose a unified formulation (Sec 3.1). We further propose to add deep fusion between image and text, making the detection model language-aware and thus a strong grounding model (Sec 3.2). With the reformulation and deep fusion, we can pre-train GLIP on scalable and semantic-rich grounding data (Sec 3.3).

3.1. Unified Formulation

Background: object detection. A typical detection model feeds an input image into a visual encoder Enc_I, with CNN [15, 51] or Transformer [34, 60, 62] as backbone, and extracts region/box features O, as shown in Figure 1 (Bottom). Each region/box feature is fed into two prediction heads, *i.e.*, a box classifier $\mathcal{C}$ and a box regressor $\mathcal{R}$, which are trained with the classification loss $\mathcal{L}_{\text{cls}}$ and the localization loss $\mathcal{L}_{\text{loc}}$, respectively:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{loc}}. \tag{1}$$

In two-stage detectors, a separate region proposal network (RPN) with RPN loss $\mathcal{L}_{\text{rpn}}$ is used to distinguish foreground from background and refine anchors. Since $\mathcal{L}_{\text{rpn}}$ does not use semantic information of object classes, we merge it into the localization loss $\mathcal{L}_{\text{loc}}$. In one-stage detectors, localization loss $\mathcal{L}_{\text{loc}}$ may also contain the centerness loss [52].

The box classifier $\mathcal{C}$ is typically a simple linear layer, and the classification loss $\mathcal{L}_{\text{cls}}$ can be written as:

$$O = \text{Enc}_I(\text{Img}),\ S_{\text{cls}} = OW^T,\ \mathcal{L}_{\text{cls}} = loss(S_{\text{cls}}; T). \tag{2}$$

Here[3], $O \in \mathbb{R}^{N \times d}$ are the object/region/box features of the input image, $W \in \mathbb{R}^{c \times d}$ is the weight matrix of the box classifier $\mathcal{C}$, $S_{\text{cls}} \in \mathbb{R}^{N \times c}$ are the output classification logits, $T \in \{0, 1\}^{N \times c}$ is the target matching between regions and classes computed from the classical many-to-1 matching [8, 31, 43, 44] or the bipartite Hungarian match [4, 9, 67]. $loss(S; T)$ is typically a cross-entropy loss for two-stage detectors and a focal loss [31] for one-stage detectors.

Object detection as phrase grounding. Instead of classifying each region/box into c classes, we reformulate detection as a grounding task, by grounding/aligning each region to c phrases in a text prompt (see Figure 1). How to design a text prompt for a detection task? Given object classes [person, bicycle, car, ..., toothbrush], one simple way is

$$\text{Prompt} = \text{“Detect: person, bicycle, car, ... , toothbrush”},$$

in which each class name is a candidate phrase to be grounded. One could design better prompts, by providing more expressive descriptions of these classes and/or by exploiting the preference of a pre-trained language model. For example, when the pre-trained BERT model [10] is used to initialize our language encoder Enc_L, the prompt “person. bicycle. car. ... . toothbrush” works better than the more human-friendly prompt described above. We will discuss the prompt design in Section 5.2.

In a grounding model, we compute the alignment scores S_{ground} between image regions and words in the prompt:

$$O = \text{Enc}_I(\text{Img}),\ P = \text{Enc}_L(\text{Prompt}),\ S_{\text{ground}} = OP^\top, \tag{3}$$

[3] N is the number of region/box features, d is the visual feature hidden dimension, c is the number of object classes, and we ignore the bias in the box classifier for simplicity.

where $P \in \mathbb{R}^{M\times d}$ is the contextual word/token features from the language encoder and plays a similar role to the weight matrix W in (2), as shown in Figure 1 (Right). The grounding model, consisting of both the image encoder Enc_I and the language encoder Enc_L, is trained end-to-end by minimizing the loss defined in (1) & (2), with a simple replacement of the classification logits S_{cls} in (2) with the region-word aligment scores S_{ground} in (3).

However, in (2), we now have the logits $S_{\text{ground}} \in \mathbb{R}^{N\times M}$ and the target $T \in \{0,1\}^{N\times c}$. The number of (sub)-word tokens M is always larger than the number of phrases c in the text prompt due to four reasons: 1) some phrases contain multiple words, e.g., "traffic light"; 2) some single-word phrases are splitted into multiple (sub)-word tokens, e.g., "toothbrush" to "tooth#" and "#brush"; 3) some are the added tokens, such as "Detect:", ",", special tokens in language models, and 4) a [NoObj] token is added at the end of the tokenized sequence. When the $loss$ is a (focal) binary sigmoid loss (the $loss$ we use in Section 4 & 5), we expand the original target matrix $T \in \{0,1\}^{N\times c}$ to $T' \in \{0,1\}^{N\times M}$ by making all sub-words positive match if a phrase is a positive match and all added tokens negative match to all image features. With this change, the $loss(S_{\text{ground}}; T')$ remains the same. During inference, we average token probabilities as the phrase probability. [4]

Equivalence between detection and grounding. With the above reformulation, we can convert any detection model into a grounding model, and the two views, i.e., detection and grounding, are theoretically equivalent for both training and inference. We also verify this empirically: the SoTA DyHead detector [9] with Swin-Tiny backbone gives the same performance on COCO val2017 before and after our reformulation. Please refer to the appendix for discussions. With the reformulation, a pre-trained phrase grounding model can be directly applied to any object detection task, thanks to the free-form input of the language encoder. This makes it possible to transfer our GLIP model to arbitrary detection tasks in a zero-shot manner.

Related work. Our grounding formulation is inspired by MDETR [19], and our grounding loss shares the same spirit of MDETR's fine-grained contrastive loss. We go further than MDETR by finding an effective approach to reformulate detection as grounding and a simple unified loss for both detection and grounding tasks. Our grounding model also resembles models for zero-shot detection [1, 12, 41, 42, 66]. The seminal work of Bansal et al. [1] enables a detection model to conduct zero-shot detection, by using the pre-trained Glove word embedding [38] as the phrase features $P \in \mathbb{R}^{c\times d}$, if written in the form of (3). Recently, phrase features extracted from pre-trained deep language models are introduced in open-vocabulary detection [61]. GLIP differs from zero-shot detection in that GLIP provides a unified view of detection and grounding, and enables two crucial ingredients, *i.e.*, language-aware deep fusion and scaling up with image-text data, as to be described next.

[4] When the $loss$ is a multi-class cross entropy (CE) loss, following MDETR [19], all box proposals with no positive match are matched to the [NoObj] token. The $loss(S, T')$ becomes a multi-label multi-class CE loss, and we sum token probabilities as phrase probability during inference.

3.2. Language-Aware Deep Fusion

In (3), the image and text are encoded by separate encoders and only fused at the end to calculate the alignment scores. We call such models *late-fusion* models. In vision-language literature [7, 19, 27, 28, 30, 36, 48, 50, 65], deep fusion of visual and language features is necessary to learn a performant phrase grounding model. We introduce deep fusion between the image and language encoders, which fuses the image and text information in the last few encoding layers, as shown in Figure 1 (Middle). Concretely, when we use DyHead [9] as the image encoder and BERT [10] as the text encoder, the deep-fused encoder is:

$$O^i_{\text{t2i}}, P^i_{\text{i2t}} = \text{X-MHA}(O^i, P^i), \quad i \in \{0, 1, .., L-1\} \tag{4}$$

$$O^{i+1} = \text{DyHeadModule}(O^i + O^i_{\text{t2i}}), \quad O = O^L, \tag{5}$$

$$P^{i+1} = \text{BERTLayer}(P^i + P^i_{\text{i2t}}), \quad P = P^L, \tag{6}$$

where L is the number of DyHeadModules in DyHead [9], BERTLayer is *newly-added* BERT Layers on top of the pre-trained BERT, O^0 denote the visual features from the vision backbone, and P^0 denote the token features from the language backbone (BERT). The cross-modality communication is achieved by the cross-modality multi-head attention module (X-MHA) (4), followed by the single modality fusion and updated in (5) & (6). Without added context vectors (O^i_{t2i} for vision modality and P^i_{i2t} for language modality), the model is reduced to a *late-fusion* model.

In the cross-modality multi-head attention module (X-MHA) (4), each head computes the context vectors of one modality by attending to the other modality:

$$O^{(q)} = OW^{(q,I)}, P^{(q)} = PW^{(q,L)}, Attn = O^{(q)}(P^{(q)})^\top/\sqrt{d},$$
$$P^{(v)} = PW^{(v,L)}, \quad O_{\text{t2i}} = \text{SoftMax}(Attn)P^{(v)}W^{(out,I)},$$
$$O^{(v)} = OW^{(v,I)}, \quad P_{\text{i2t}} = \text{SoftMax}(Attn^\top)O^{(v)}W^{(out,L)},$$

where $\{W^{(\text{symbol},I)}, W^{(\text{symbol},L)} : \text{symbol} \in \{q, v, out\}\}$ are trainable parameters and play similar roles to those of query, value, and output linear layers in Multi-Head Self-Attention [53], respectively.

The deep-fused encoder (4)-(6) brings two benefits. 1) It improves the phrase grounding performance. 2) It makes the learned visual features language-aware, and thus the model's prediction is conditioned on the text prompt. This is crucial to achieve the goal of having one model serve all downstream detection tasks (shown in Section 5.2).

3.3. Pre-training with Scalable Semantic-Rich Data

Considerable efforts have been devoted to collecting detection data that are rich in semantics and large in quantity. However, human annotations have been proven costy and limited [13, 25]. Prior work seeks to scale up in a self-training fashion [68]. They use a teacher (a pre-trained detector) to predict boxes from raw images and generate pseudo detection labels to train a student model. But the generated data are still limited in terms of the size of the concept pool, as the teacher can only predict labels defined in the concept pool, constructed on the existing datasets. In contrast, our model can be trained on both detection and, more importantly, grounding data. We show that grounding data can provide rich semantics to facilitate localization and can be scaled up in a self-training fashion.

First, the gold grounding data cover a much larger vocabulary of visual concepts than existing detection data. The largest attempts at scaling up detection vocabulary still cover no more than 2,000 categories [13, 23]. With grounding data, we expand the vocabulary to cover virtually any concepts that appear in the grounded captions. For example, Flickr30K [39] contains 44,518 unique phrases while VG Caption [23] contains 110,689 unique phrases, orders of magnitude larger than the vocabulary of detection data. We provide an empirical study in Section 4.4 to show that 0.8M gold grounding data brings a larger improvement on detecting rare categories than additional 2M detection data.

Further, instead of scaling up detection data, we show a promising route to obtaining semantically rich data: scaling up grounding data. We use a simple approach inspired by self-training. We first pre-train a *teacher* GLIP with gold (human-annotated) detection and grounding data. Then we use this teacher model to predict boxes for web-collected image-text data, with noun phrases detected by an NLP parser [2]. Finally, a *student* model is trained with both the gold data and the generated pseudo grounding data. As shown in Figure 2, the teacher is capable of generating accurate boxes for semantically rich entities.

Why can the student model possibly outperform the teacher model? While discussions remain active in the self-training literature [68], in the context of visual grounding, we posit that the teacher model is utilizing the language context and language generalization ability to accurately ground concepts that it may not inherently know. For example, in Figure 2, the teacher may not directly recognize certain concepts such as *vaccine* and *turquoise*, if they are not present in gold data. However, the rich language context such as syntactic structures can provide strong guidance for the teacher model to perform an "educated guess". The model can localize *vaccine* if it can localize *a small vail*; it can localize *turquoise* if it can find *caribbean sea*. When we train the student model, the "educated guess" of the teacher model becomes a "supervised signal", enabling the student model to learn the concept of *vaccine* and *turquoise*.

4. Transfer to Established Benchmarks

After pre-training, GLIP can be applied to grounding and detection tasks with ease. We show strong direct domain transfer performance on three established benchmarks: 1) MS-COCO object detection (COCO) [32] containing 80 common object categories; 2) LVIS [13] covering over 1000 objects categories; 3) Flickr30K [39], for phrase grounding. We train 5 variants of GLIP (Table 1) to ablate its three core techniques: 1) unified grounding loss; 2) language-aware deep fusion; 3) and pre-training with both types of data. Implementation deails are in the appendix.

GLIP-T (A) is based on a SoTA detection model, Dynamic Head [9], with our word-region alignment loss replacing the classification loss. It is based on the Swin-Tiny backbone and pre-trained on O365 (Objects365 [45]), which contains 0.66M images and 365 categories. As discussed in Section 3.1, the model can be viewed as a strong classical zero-shot detection model [1], relying purely on the language encoder to generalize to new concepts.

GLIP-T (B) is enhanced with language-aware deep fusion but pre-trained only on O365.

GLIP-T (C) is pre-trained on 1) O365 and 2) GoldG, 0.8M human-annotated gold grounding data curated by MDETR [19], including Flickr30K, VG Caption [23], and GQA [16]. We have removed COCO images from the dataset. It is designed to verify the effectiveness of gold grounding data

GLIP-T is based on the Swin-Tiny backbone and pre-trained on the following data: 1) O365, 2) GoldG as in GLIP-T (C), and 3) Cap4M, 4M image-text pairs collected from the web with boxes generated by GLIP-T (C). We also experiment with existing image caption datasets: CC (Conceptual Captions with 3M data) [46] and SBU (with 1M data) [37]. We find that CC+SBU GLIP-T performs slightly better than Cap4M GLIP-T on COCO, but slightly worse on the other datasets. For simplicity, we report both versions on COCO but only the Cap4M model for the other tasks. We present the full results in the appendix.

GLIP-L is based on Swin-Large and trained with: 1) FourODs (2.66M data), 4 detection datasets including Objects365, OpenImages [22], Visual Genome (excluding COCO images) [23], and ImageNetBoxes [24]; 2) GoldG as in GLIP-T (C); and 3) CC12M+SBU, 24M image-text data collected from the web with generated boxes.

4.1. Zero-Shot and Supervised Transfer on COCO

We conduct experiments on MS-COCO to evaluate models' transfer ability to common categories. We evaluate under two settings: 1) zero-shot domain transfer, and 2) supervised transfer, where we fine-tune the pre-trained models using the standard setting. For the fine-tuning setting, we additionally test the performance of a GLIP-L model, where

Model	Backbone	Deep Fusion	Pre-Train Data		
			Detection	Grounding	Caption
GLIP-T (A)	Swin-T	✗	Objects365	-	-
GLIP-T (B)	Swin-T	✓	Objects365	-	-
GLIP-T (C)	Swin-T	✓	Objects365	GoldG	-
GLIP-T	Swin-T	✓	Objects365	GoldG	Cap4M
GLIP-L	Swin-L	✓	FourODs	GoldG	Cap24M

Table 1. A detailed list of GLIP model variants.

Model	Backbone	Pre-Train Data	Zero-Shot	Fine-Tune
			2017val	2017val / test-dev
Faster RCNN	RN50-FPN	-	-	40.2 / -
Faster RCNN	RN101-FPN	-	-	42.0 / -
DyHead-T [9]	Swin-T	-	-	49.7 / -
DyHead-L [9]	Swin-L	-	-	58.4 / 58.7
DyHead-L [9]	Swin-L	O365,ImageNet21K	-	60.3 / 60.6
SoftTeacher [58]	Swin-L	O365,SS-COCO	-	60.7 / 61.3
DyHead-T	Swin-T	O365	43.6	53.3 / -
GLIP-T (A)	Swin-T	O365	42.9	52.9 / -
GLIP-T (B)	Swin-T	O365	44.9	53.8 / -
GLIP-T (C)	Swin-T	O365,GoldG	**46.7**	55.1 / -
GLIP-T	Swin-T	O365,GoldG,Cap4M	46.3	54.9 / -
GLIP-T	Swin-T	O365,GoldG,CC3M,SBU	46.6	**55.2** / -
GLIP-L	Swin-L	FourODs,GoldG,Cap24M	**49.8**	**60.8** / 61.0
GLIP-L	Swin-L	FourODs,GoldG+,COCO	-	- / **61.5**

Table 2. Zero-shot domain transfer and fine-tuning on COCO. GLIP, without seeing any images from the COCO dataset, can achieve comparable or superior performance than prior supervised models (e.g. GLIP-T under Zero-Shot v.s. Faster RCNN under Fine-Tune). When fully fine-tuned on COCO, GLIP-L surpasses the SoTA performance.

we include the COCO images in the pre-training data (the last row). Specifically, we add the full GoldG+ grounding data and COCO train2017 to the pre-training data. Note that part of COCO 2017val images are present in GoldG+ [19]. Thus we only report the test-dev performance of this model. Please see more details in the appendix.

We introduce an additional baseline: DyHead pre-trained on Objects365. We find that COCO 80 categories are fully covered in Objects365. Thus we can evaluate DyHead trained on Objects365 in a "zero-shot" way: during inference, instead of predicting from 365 classes, we restrict the model to predict only from the COCO 80 classes. We list standard COCO detection models for reference. We also list two state-of-the-art models pre-trained with extra data.

Results are present in Table 2. Overall, GLIP models achieve strong zero-shot and supervised performance. **Zero-shot GLIP models rival or surpass well-established supervised models**. The best GLIP-T achieves 46.7 AP, surpassing Faster RCNN; GLIP-L achieves 49.8 AP, surpassing DyHead-T. Under the supervised setting, the best GLIP-T brings 5.5 AP improvement upon the standard DyHead (55.2 v.s. 49.7). With the Swin-Large backbone, **GLIP-L surpasses the current SoTA on COCO**, reaching 60.8 on 2017val and 61.5 on test-dev, without some bells and whistles in prior SoTA [58] such as model EMA, mix-up, label smoothing, or soft-NMS.

Model	Backbone	MiniVal [19]				Val v1.0			
		APr	APc	APf	AP	APr	APc	APf	AP
MDETR [19]	RN101	20.9	24.9	24.3	24.2	-	-	-	-
MaskRCNN [19]	RN101	26.3	34.0	33.9	33.3	-	-	-	-
Supervised-RFS [13]	RN50	-	-	-	-	12.3	24.3	32.4	25.4
GLIP-T (A)	Swin-T	14.2	13.9	23.4	18.5	6.0	8.0	19.4	12.3
GLIP-T (B)	Swin-T	13.5	12.8	22.2	17.8	4.2	7.6	18.6	11.3
GLIP-T (C)	Swin-T	17.7	19.5	**31.0**	24.9	7.5	11.6	**26.1**	16.5
GLIP-T	Swin-T	**20.8**	**21.4**	**31.0**	**26.0**	**10.1**	**12.5**	25.5	**17.2**
GLIP-L	Swin-L	**28.2**	**34.3**	**41.5**	**37.3**	**17.1**	**23.3**	**35.4**	**26.9**

Table 3. Zero-shot domain transfer to LVIS. While using no LVIS data, GLIP-T/L outperforms strong supervised baselines (shown in gray). Grounding data (both gold and self-supervised) bring large improvements on APr.

Row	Model	Data	Val			Test		
			R@1	R@5	R@10	R@1	R@5	R@10
1	MDETR-RN101	GoldG+	82.5	92.9	94.9	83.4	93.5	95.3
2	MDETR-ENB5	GoldG+	83.6	93.4	95.1	84.3	93.9	95.8
3	GLIP-T	GoldG	84.0	95.1	96.8	84.4	95.3	97.0
4		O365,GoldG	84.8	94.9	96.3	85.5	95.4	96.6
5		O365,GoldG,Cap4M	**85.7**	**95.4**	**96.9**	**85.7**	**95.8**	**97.2**
6	GLIP-L	FourODs,GoldG,Cap24M	**86.7**	**96.4**	**97.9**	**87.1**	**96.9**	**98.1**

Table 4. Phrase grounding performance on Flickr30K entities. GLIP-L outperforms previous SoTA by 2.8 points on test R@1.

4.2. Zero-Shot Transfer on LVIS

We evaluate the model's ability to recognize diverse and rare objects on LVIS in a zero-shot setting. We report on MiniVal containing 5,000 images introduced in MDETR as well as the full validation set v1.0. Please see the evaluation details in the appendix.

Results are present in Table 3. We list three supervised models trained on the annotated data of LVIS. GLIP exhibits strong zero-shot performance on all the categories. **GLIP-T is on par with supervised MDETR while GLIP-L outperforms Supervised-RFS by a large margin.**

The benefit of using grounding data is evident. Gold grounding data brings a 4.2-point improvement on MiniVal APr (model C v.s. model B). Adding image-text data further improves performance by 3.1 points. We conclude that the semantic richness of grounding data significantly helps the model recognize rare objects.

4.3. Phrase Grounding on Flickr30K Entities

We evaluate the model's ability to ground entities in natural language on Flickr30K entities [39]. Flickr30K is included in the gold grounding data so we directly evaluate the models after pre-training as in MDETR [19]. We use the any-box-protocol specified in MDETR. Results are present in Table 4. We evaluate three versions of GLIP with different pre-training data. We list the performance of MDETR, the SoTA grounding model. MDETR is trained on GoldG+, containing 1.3M data (GoldG is a subset of GoldG+ excluding COCO images).

GLIP-T with GoldG (Row 3) achieves similar perfor-

Row	Pre-Training Data	COCO 2017val	LVIS MiniVal			
			AP_r	AP_c	AP_f	AP
1	VG w/o COCO	26.9	4.9	10.4	23.2	16.1
2	+ GoldG	29.2	7.8	14.0	24.5	18.5
3	OpenImages	29.9	12.8	12.1	17.8	14.9
4	+ GoldG	33.6	15.2	16.9	24.5	20.4
5	O365	44.9	13.5	12.8	22.2	17.8
6	+GoldG	**46.7**	17.7	19.5	31.0	24.9
7	O365,GoldG,Cap4M	46.3	**20.8**	21.4	31.0	26.0
8	FourODs	46.3	15.0	**22.5**	**32.8**	**26.8**

Table 5. Effect of different detection data.

mance to MDETR with GoldG+, presumably due to the introduction of Swin Transformer, DyHead module, and deep fusion. More interestingly, the addition of detection data helps grounding (Row 4 v.s. 3), showing again the synergy between the two tasks and the effectiveness of our unified loss. Image-text data also helps (Row 5 v.s. 4). Lastly, scaling up (GLIP-L) can achieve 87.1 Recall@1, outperforming the previous SoTA by 2.8 points.

4.4. Analysis

In this section, we perform ablation study by pre-training GLIP-T on different data sources (Table 5). We answer two research questions. First, our approach assumes the use of a detection dataset to bootstraps the model. One natural question is whether grounding data brings improvement when paired with different detection data. We find that adding grounding data brings consistent improvement with different detection data (Row 1-6).

Second, we have shown the effectiveness of grounding data for both common and rare categories. One orthogonal direction is to scale up detection data by including more images and categories (Section 3.3). We intend to provide an empirical comparison between scaling up detection data and grounding data. We present GLIP trained with 4 public detection datasets (Row 8) as an extreme attempt at scaling up detection data with human annotations. The model is trained with 2.66M detection data in total, with an aligned vocabulary of over 1,500 categories. However, it still trails behind Row 6 on COCO and AP_r of LVIS, where Row 6 is trained with only 0.66M detection data and 0.8M gold grounding data. Adding image-text data further widens the gap on LVIS AP_r (20.8 versus 15.0). We conclude that grounding data are indeed more semantic-rich and a promising alternative to scaling up detection data.

5. Object Detection in the Wild

To evaluate GLIP's transferability to diverse real-world tasks, we curate an "Object Detection in the Wild" (ODinW) setting. We choose 13 public datasets on Roboflow[5], each requiring a different localization skill.

[5] https://public.roboflow.com/object-detection

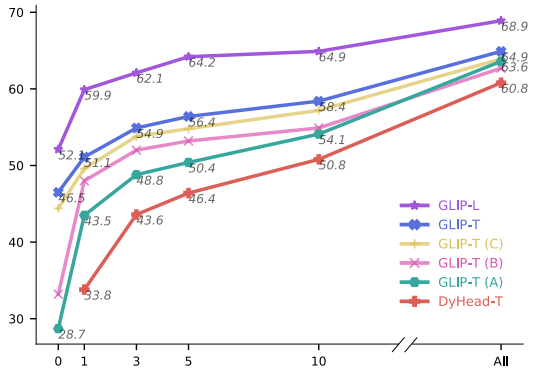

Figure 3. Data efficiency of models. X-axis is the amount of task-specific data, from zero-shot to all data. Y-axis is the average AP across 13 datasets. GLIP exhibits great data efficiency, while each of our proposed approach contributes to the data efficiency.

Many of the datasets are designed with a specific application purpose to mimic real-world deployment scenarios. For example, EgoHands requires locating hands of a person; Pothole concerns detecting holes on the road; ThermalDogsandPeople involves identifying dogs and persons in infrared images. Please refer to the appendix for details.

We demonstrate that GLIP facilitates transfer to such diverse tasks. (1) GLIP brings great data efficiency, reaching the same performance with significantly less task-specific data than baselines (Section 5.1). (2) GLIP enables new domain transfer strategies: when adapting to a new task, we can simply change the text prompt and keep the entire grounding model unchanged. This greatly reduces deployment cost because it allows one centralized model to serve various downstream tasks (Section 5.2).

5.1. Data Efficiency

We vary the amount of task-specific annotated data, from zero-shot (no data provided), to X-shot (providing at least X examples per category [20, 55, 59]), to using all data in the training set. We fine-tune the models on the provided data and use the same hyper-parameters for all models. Each dataset comes with pre-specified category names. As GLIP is language-aware, we find it beneficial to re-write some pre-specified names with more descriptive language (see Section 5.2 for a discussion). We compare with the SoTA detector DyHead-T, pre-trained on Objects365. We test with the standard COCO-trained DyHead-T and find it giving similar performance. For simplicity, we report only the former. We also experiment with the scaled cosine similarity approach [54] but find it slightly underperforming the vanilla approach so we report only the latter. Please refer to the appendix for full statistics, including three independent runs for X-shot experiments.

Results are shown in Figure 3. We find that unified grounding reformulation, deep fusion, grounding data, and

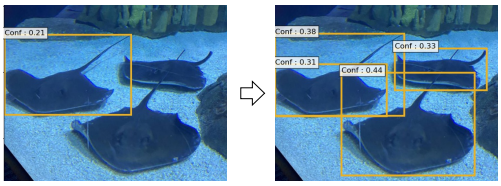

Figure 4. A manual prompt tuning example from the Aquarium dataset in ODinW. Given an expressive prompt ("flat and round"), zero-shot GLIP can detect the novel entity "stingray" better.

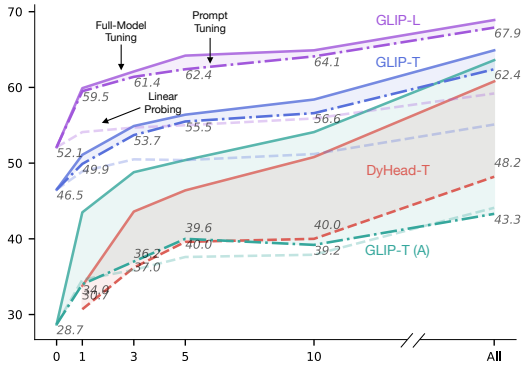

Figure 5. Effectiveness of prompt tuning. Solid lines are full-model tuning performance; dashed lines are prompt/linear probing performance. By only tuning the prompt embeddings, GLIP-T and GLIP-L can achieve performance close to full-model tuning, allowing for efficient deployment.

model scale-up all contribute to the improved data efficiency (from the bottom red line (Dyhead-T) up to the upper purple line (GLIP-L)). As a result, GLIP exhibits transformative data efficiency. A zero-shot GLIP-T outperforms 5-shot DyHead-T while a one-shot GLIP-L is competitive with a fully supervised DyHead-T.

5.2. One Model for All Tasks

As neural models become larger, how to reduce deployment cost has drawn an growing research interest. Recent work on language models [47], image classification [64], and object detection [54] has explored adapting a pre-trained model to a new domain but only changing the least amount of parameters. Such a setting is often denoted as linear probing [21], prompt tuning [64], or efficient task adapters [11]. The goal is to have a single model serving various tasks, and each task adds only a few task-specific parameters or no parameters to the pre-trained model. This reduces training and storage cost. In this section, we evaluate models against the metric of deployment efficiency.

Manual prompt tuning. As GLIP performs language-aware localization, i.e., the output of GLIP is heavily conditioned on the language input, we propose an efficient way for GLIP to do task transfer: for any novel categories, the user can use expressive descriptions in the text prompt, adding attributes or language context, to inject domain knowledge and help GLIP transfer. For example, on the left hand side of Figure 4, the model fails to localize all occurrences of the novel entity "stingray". However, by adding the attributes to the prompt, i.e., "flat and round", the model successfully localizes all occurrences of stingrays. With this simple prompt change, we improve the AP50 on stingray from 4.6 to 9.7. This resembles the prompt design technique in GPT-3 [3] and is practically appealing, as it requires no annotated data or model re-training. Please refer to the appendix for more details.

Prompt tuning. We further consider the setting where we have access to task-specific training data but wish to tune the least amount of parameters for easy deployment. For classical detection models, Wang *et al*. [54] report the effectiveness of "linear probing" (i.e., train only the box regression and classification head). GLIP can also be "linear probed", where we only fine-tune the box head and a projection layer between the region and prompt embeddings. Because of the language-aware deep fusion, GLIP supports a more powerful yet still efficient transfer strategy: prompt tuning [26, 47]. For GLIP, as each detection task has only one language prompt (e.g., the prompt for Pothole could be "Detect pothole." for all images), we first get prompt embeddings P^0 from the language backbone, then discard the language backbone and only fine-tune P^0 as the task-specific input (Section 3.2).

We evaluate the models' performance under three settings (Figure 5): linear probing, prompt tuning (only applicable for GLIP), and full-model tuning. For DyHead-T, prompt tuning is not applicable as the traditional object detection model cannot accept language input; the gap between linear probing and full-model tuning is large. GLIP-T (A) has no language-aware deep fusion; thus prompt tuning and linear tuning achieve similar performance and lag significantly behind full-model tuning. However, for GLIP-T and GLIP-L, prompt tuning almost matches the full-tuning results, without changing any of the grounding model parameters. Interestingly, as the model and data size grow larger, the gap between full-model tuning and prompt tuning becomes smaller (GLIP-L v.s. GLIP-T), echoing the findings in NLP literature [33].

6. Conclusion

GLIP unifies the object detection and phrase grounding tasks to learn an object-level, language-aware, and semantic-rich visual representation. After pre-training, GLIP showed promising results on zero-shot and fine-tuning settings on well-established benchmarks and 13 downstream tasks. We leave a detailed study of how GLIP scales with text-image data size to future work.

References

[1] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 384–400, 2018. 3, 4, 5

[2] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009. 2, 5

[3] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 8

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 3

[5] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4983, 2019. 3

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 1

[7] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*, 2019. 1, 4

[8] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016. 3

[9] Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. Dynamic head: Unifying object detection heads with attentions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7373–7382, 2021. 2, 3, 4, 5, 6

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3, 4

[11] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021. 8

[12] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Zero-shot detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 3, 4

[13] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5356–5364, 2019. 5, 6

[14] Xiaotian Han, Jianwei Yang, Houdong Hu, Lei Zhang, Jianfeng Gao, and Pengchuan Zhang. Image scene graph generation (sgg) benchmark, 2021. 1

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[16] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 5

[17] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as composition of spatio-temporal scene graphs, 2019. 1

[18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning (ICML)*, 2021. 3

[19] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 3, 4, 5, 6

[20] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8420–8429, 2019. 7

[21] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018. 8

[22] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2(3):18, 2017. 5

[23] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*, 123(1):32–73, 2017. 2, 5

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 5

[25] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*, 2018. 1, 2, 5

[26] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. 8

[27] Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training. *arXiv preprint arXiv:1908.06066*, 2019. 1, 4

[28] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 1, 4

[29] Liunian Harold Li, Haoxuan You, Zhecan Wang, Alireza Zareian, Shih-Fu Chang, and Kai-Wei Chang. Unsupervised vision-and-language pre-training without parallel images and captions. *arXiv preprint arXiv:2010.12831*, 2020. 1

[30] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 121–137. Springer, 2020. 1, 4

[31] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 1, 3

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 2, 5

[33] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*, 2021. 8

[34] 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. *arXiv preprint arXiv:2103.14030*, 2021. 3

[35] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[36] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 13–23, 2019. 1, 4

[37] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24:1143–1151, 2011. 5

[38] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics, 2014. 4

[39] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 5, 6

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. 1, 3

[41] Shafin Rahman, Salman Khan, and Nick Barnes. Improved visual-semantic alignment for zero-shot object detection. In *34th AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 3, 4

[42] Shafin Rahman, Salman H Khan, and Fatih Porikli. Zero-shot object detection: Joint recognition and localization of novel concepts. *International Journal of Computer Vision*, 128(12):2979–2999, 2020. 3, 4

[43] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 3

[44] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015. 1, 3

[45] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 8430–8439, 2019. 2, 5

[46] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2556–2565, 2018. 5

[47] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*, 2020. 8

[48] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 1, 4

[49] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019. 1

[50] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 1, 4

[51] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International*

Conference on Machine Learning, pages 6105–6114. PMLR, 2019. 3

[52] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 3

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 4

[54] Xin Wang, Thomas E Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. *arXiv preprint arXiv:2003.06957*, 2020. 7, 8

[55] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Meta-learning to detect rare objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9925–9934, 2019. 7

[56] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018. 1

[57] Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017. 1

[58] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. *arXiv preprint arXiv:2106.09018*, 2021. 2, 6

[59] Xiaopeng Yan, Ziliang Chen, Anni Xu, Xiaoxi Wang, Xiaodan Liang, and Liang Lin. Meta r-cnn: Towards general solver for instance-level low-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9577–9586, 2019. 7

[60] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal self-attention for local-global interactions in vision transformers. *arXiv preprint arXiv:2107.00641*, 2021. 3

[61] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021. 3, 4

[62] Pengchuan Zhang, Xiyang Dai, Jianwei Yang, Bin Xiao, Lu Yuan, Lei Zhang, and Jianfeng Gao. Multi-scale vision longformer: A new vision transformer for high-resolution image encoding. *arXiv preprint arXiv:2103.15358*, 2021. 3

[63] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021. 1

[64] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *arXiv preprint arXiv:2109.01134*, 2021. 8

[65] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and VQA. *AAAI*, 2020. 1, 4

[66] Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama. Don't even look once: Synthesizing features for zero-shot detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3, 4

[67] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 3

[68] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking pre-training and self-training. *Advances in Neural Information Processing Systems*, 33, 2020. 2, 5