Cycle-Consistency Learning for Captioning and Grounding

Ning Wang[1], Jiajun Deng[2], Mingbo Jia[1]

[1]Huawei Inc.
[2]University of Adelaide, Australian Institute for Machine Learning
wn6149@mail.ustc.edu.cn, jiajun.deng@adelaide.edu.au, jiamingbo@huawei.com

Abstract

We present that visual grounding and image captioning, which perform as two mutually inverse processes, can be bridged together for collaborative training by careful designs. By consolidating this idea, we introduce CyCo, a cyclic-consistent learning framework to ameliorate the independent training pipelines of visual grounding and image captioning. The proposed framework (1) allows the semi-weakly supervised training of visual grounding; (2) improves the performance of fully supervised visual grounding; (3) yields a general captioning model that can describe arbitrary image regions. Extensive experiments show that our fully supervised grounding model achieves state-of-the-art performance, and the semi-weakly supervised one also exhibits competitive performance compared to the fully supervised counterparts. Our image captioning model has the capability to freely describe image regions and meanwhile shows impressive performance on prevalent captioning benchmarks.

Introduction

The recent decades have witnessed the great success in vision-language (VL) related fields. Based on the primary target of bridging the modality gap between vision and language, deep neural networks addressing VL tasks generally share the pre-training objective, model structure, large-scale training corpus, *etc*. However, by the time of downstream fine-tuning, these tasks are typically individually tackled or simply combined in a multi-task training paradigm.

In this work, we devote our efforts to two VL downstream tasks including image captioning (Vinyals et al. 2015; Anderson et al. 2018) and visual grounding (Mao et al. 2016; Yu et al. 2016), and explore their inherent relationships to enable effective joint training. To match the granularity of visual parts in visual grounding, we first extend image captioning to a more general scenario, where the model is intended to describe a given region. We define this generalized task as regional image captioning, which is similar to dense captioning task (Johnson, Karpathy, and Fei-Fei 2016) but is free of the requirement of object detection. Particularly, the conventional task definition of image captioning (Vinyals et al. 2015) is a special case of regional image captioning

(a) Captioning-to-Grounding Cycle-Consistency Framework

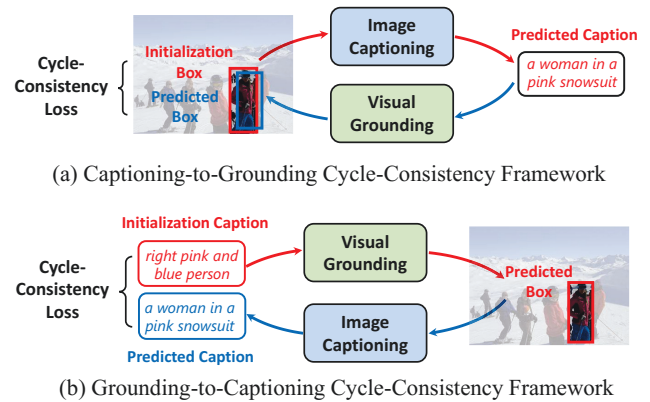(b) Grounding-to-Captioning Cycle-Consistency Framework

Figure 1: The proposed framework jointly optimizes image captioning and visual grounding models via cycle-consistency learning. In (a), our method computes the region (box) consistency in the captioning-to-grounding optimization cycle. In (b), our framework measures the caption consistency in the grounding-to-captioning cyclic process.

that regards the whole image as a region. For visual grounding, which is also known as referring expression comprehension (Mao et al. 2016; Yu et al. 2016) and phrase localization (Kazemzadeh et al. 2014; Plummer et al. 2015) in the literature, we maintain the original task target to localize the corresponding region (generally denoted by a bounding box) described by a given language expression.

Despite achieving inspiring progress, image captioning and visual grounding still suffer from several limitations. For visual grounding, this task simultaneously requires text descriptions and accurate object bounding boxes for model optimization. These fine-grained image-text-box triplets are rather laborious to collect. How to optimize the grounding model using limited data annotations has received considerable attention (Liu et al. 2021; Wang et al. 2021a). As for image captioning, existing methods typically focus on describing the whole image. The capabilities of modeling region relationships and properly describing them are largely overlooked in existing algorithms. We argue that a robust image captioner should be qualified to freely describe the image, from an arbitrary region to the whole image. Simi-

lar to the grounding task, obtaining the regional captioning ability also requires sufficient image-text-box data, increasing the training cost.

In this paper, we introduce a joint learning framework, namely CyCo, to ameliorate the training pipelines of image captioning and visual grounding via cyclic consistency. Our core motivation is that visual grounding and regional image captioning can be regarded as an inverse process of each other. Specifically, a visual grounding model takes an image and a region-specific description as inputs to predict the position of the corresponding bounding box, while regional image captioning receives the location of a region to produce a region-aware caption. When taking one's output as the input of the other, these two tasks naturally establish a cyclic structure. As shown in Figure 1 (a), the generated bounding box from the grounding model is expected to be consistent with the input box of the regional image captioner. In this process, the whole framework merely needs an initialized bounding box for training. As depicted in Figure 1 (b), the produced text description from the regional image captioner is expected to be consistent with the input referring text of the grounding model. As a result, the models in this cycle are free of the bounding box annotations.

The proposed cycle-consistency learning framework enjoys the following merits. (1) We bridge two independent VL tasks in a unified framework. To this end, our method can potentially absorb the training data from both tasks and share the model parameters for collaborative training. (2) Thanks to the joint training and online data augmentation (iterative pseudo label generation) in the cyclic learning process, under the *same* training data, our framework further improves the performance of the supervised grounding model. (3) After collaborative training, our framework yields a strong image captioning model that can describe the visual contents in different spatial levels, *i.e.*, from a subregion to the global image. (4) The proposed framework allows the semi-weakly supervised training of the grounding model, which merely needs limited fully-annotated images and many more images with only bounding box annotations *or* only language expression labels for model training.

In summary, we make three-fold contributions:

- We present a novel cycle-consistency learning framework to bridge two independent vision-language tasks including image captioning and visual grounding.
- We design a simple regional image captioner and a Transformer-based grounding model. We further organize them in a unified framework with weight-sharing architecture for efficient end-to-end learning.
- Extensive experiments validate the effectiveness of our proposed cycle-consistency learning framework.

Related Work

Vision-language Pre-training. Vision-language (VL) pre-training algorithms (Radford et al. 2021; Li et al. 2020, 2022) aim to bridge the domain gap between vision and language representations. The recent dual-encoder methods such as CLIP (Radford et al. 2021) align the representations using contrastive learning. Despite the outstanding performance, their light interaction manner fails to deeply fuse VL representations for generation tasks. In contrast, recent VL pre-training approaches (Zhou et al. 2020; Li et al. 2020, 2021, 2022) adopt a relatively heavy Transformer architecture (Vaswani et al. 2017) to achieve the deeper multi-modal interaction. Inspired by the success of previous arts, we also conduct the cross-modal pre-training to prompt the downstream VL tasks.

Visual Grounding. Traditional visual grounding methods typically follow a two-stage pipeline, which generates plentiful region proposals in the first stage and selects the most matched one via language expression in the second stage (Yang, Li, and Yu 2019; Liu et al. 2019). Recently, one-stage visual grounding approaches gain increasing attention. They generally embed the linguistic information into the one-stage object detector (Yang et al. 2019) or model multi-modal representations via Transformer (Deng et al. 2021, 2023) for efficient visual grounding.

Different from the above algorithms based on supervised training, weakly supervised grounding models learn the region-phrase correspondence with only language expressions. These methods first obtain a set of ROIs (Region of Interest) using object detectors, and then mine the correspondence between ROIs and query expressions for model training (Wang et al. 2021a; Liu et al. 2021). In this work, we train the grounding model in a semi-weakly supervised manner with the help of a captioning model. Our framework is free of external object detectors for data pre-processing.

Image Captioning. Image captioning aims to generate a human-readable sentence to describe the image contents. Captioning algorithms (Huang et al. 2019; Anderson et al. 2018; Hu et al. 2021b) typically utilize the object detectors to extract ROI features or simply exploit the grid features for efficient visual representation modeling (Wang et al. 2021b; Li et al. 2022). After visual feature extraction, captioning models utilize a decoder such as Transformer to generate the sentence (Wang, Xu, and Sun 2022; Wang et al. 2023a,b). Previous algorithms exploit the region information to facilitate the image captioning (Chen et al. 2020a; Kinghorn, Zhang, and Shao 2018; Cornia, Baraldi, and Cucchiara 2019), but they still focus on describing the *global* image content. The dense captioner (Johnson, Karpathy, and Fei-Fei 2016) mainly focuses on detecting and describing *local* ROI regions. In contrast, our image captioner is designed to freely describe *both* global and regional contexts.

Recently, some methods (Wang et al. 2022; Yu et al. 2017) jointly train the captioning and grounding models in a multi-task fashion. Mao *et al.* (Mao et al. 2016) utilize the grounding model as a verification module to push the generated referring expressions to be unambiguous. Different from previous arts, our method bridges the captioning and grounding models in a *cyclic framework* and explores *two* different cycle-consistency constraints for collaborative training.

Cycle-Consistency Learning. To bridge one modality or domain to the other, cycle-consistency learning has been explored extensively in visual object tracking (Wang et al. 2019; Wang, Jabri, and Efros 2019), machine translation (He et al. 2016), unpaired image-to-image translation (Zhu et al. 2017), visual question answering (Shah et al. 2019), image captioning (Guo et al. 2019), *etc*. Different from the previ-

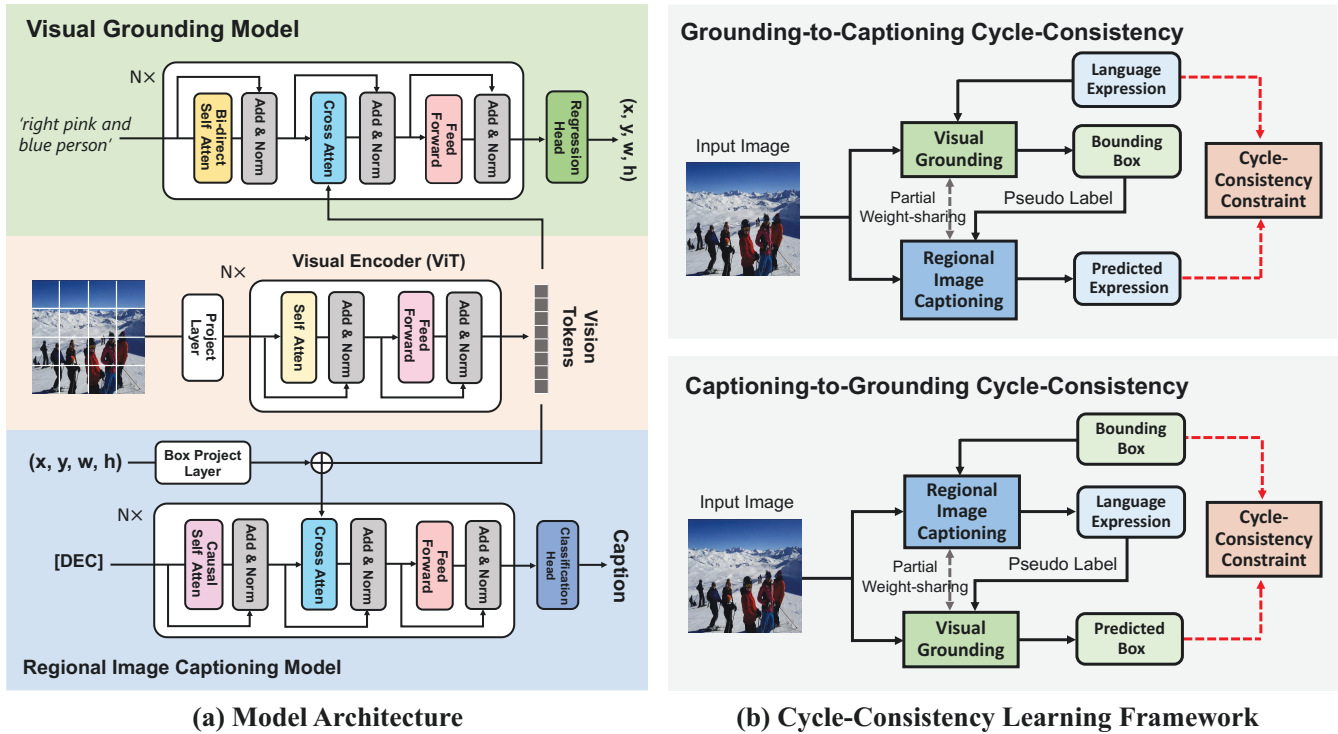**(a) Model Architecture**   **(b) Cycle-Consistency Learning Framework**

Figure 2: In (a), we exhibit the model architecture of our proposed joint training framework of visual grounding and regional image captioning. Two models share a visual encoder (ViT) and leverage different Transformer blocks for individual tasks. In (b), we show the cyclic consistency learning processes of our framework including a grounding-to-captioning cycle and a captioning-to-grounding cycle.

ous arts that explore the consistency within a single modality or a single task, we jointly optimize two VL tasks to form different cycles. To our knowledge, the cyclic consistency between image captioning and visual grounding has rarely been touched in the literature. Further, we explore the potential of our framework in different training scenarios including semi-weakly supervised and fully supervised training.

Methodology

Our method follows the pretrain-and-finetune paradigm. At the pre-training stage, we leverage the widely-adopted training objectives (*i.e.,* image-text contrastive loss, matching loss, and language modeling loss) to align and fuse the visual and linguistic representations. At the fine-tuning stage, both visual grounding and image captioning models reuse the pre-trained model architecture, while capitalizing on task-specific head networks. During fine-tuning, we further develop the cycle-consistency learning framework to jointly optimize the grounding and captioning models. The detailed model architecture and our proposed cycle-consistency learning framework are illustrated in Figure 2.

Revisiting Model Pre-training

Our pre-trained vision-language model follows the BLIP approach (Li et al. 2022). We briefly review its vision encoder, image-grounded text encoder, and image-grounded text decoder, which are highly related to our downstream tasks.

Vision Encoder. We exploit the commonly used ViT-B/16 network (Dosovitskiy et al. 2020) as the vision encoder. ViT-B/16 is composed of a stack of 12 Transformer encoder layers with 12 heads in each multi-head attention layer. Given an image, we first split it into small patches with equally 16×16 size, and project them into feature vectors, which are known as vision tokens. We denote the final encoded vision tokens as v.

Image-Grounded Text Encoder. In this block, we first append a special [ENC] token at the beginning of the text sequence. Then the text tokens are converted to embeddings via a word embedding layer. This text encoder leverages the bi-directional self-attention to further encode the text embeddings, and aggregate the visual information through cross-attention with vision tokens for multi-modal fusion.

The output embedding of [ENC] contains the rich multi-modal representation of the image-text pair, which is exploited to compute the image-text matching (ITM) loss. Specifically, we add a binary classification head on top of it to predict whether an image-text pair is matched.

Image-Grounded Text Decoder. This text decoder block is similar to the above image-grounded text encoder, while replacing the bi-directional self-attention with the causal self-attention to facilitate the text generation. A special [DEC] token is used as the beginning signal of a sequence.

The image-grounded text decoder is optimized by the language modeling (LM) loss, which maximizes the likelihood

of the text in an autoregressive manner.

Cycle-Consistency Learning Framework

In this section, we first introduce how to transfer the pre-trained model to downstream visual grounding and regional image captioning tasks. Then, we exhibit how to jointly optimize them by virtue of the cyclic constraints.

Visual Grounding (VG). To facilitate the following introduction of cycle-consistency learning, we depict the visual grounding model in a holistic view, which takes the visual features v and description tokens x as its inputs, and outputs a predicted bounding box b_{pred} as follows.

$$b_{pred} = \text{Model}_{\text{VG}}(v, x). \qquad (1)$$

In our framework, visual grounding model $\text{Model}_{\text{VG}}(\cdot, \cdot)$ reuses the image-grounded text encoder block, which fuses the visual and text features to predict the region localization. Following the setting in the pre-training stage, we also add a special [ENC] token at the beginning of the tokenized tokens. Then, the bi-directional self-attention encodes the text tokens and cross-attention injects visual information into the text tokens. After multi-layer visual-linguistic fusion via the Transformer structure, the output embedding of the [ENC] token contains rich cross-modal contexts, which is leveraged for bounding box regression via a regression head. The regression head is a simple three-layer multi-layer perceptron (MLP) with ReLU activations between layers. The output of the box regression head is the 4-dim box coordinates $b_{pred} = (x, y, w, h)$.

In the model fine-tuning stage, we normalize the ground-truth region bounding box by the scale of the image to obtain $b_{gt} = (\hat{x}, \hat{y}, \hat{w}, \hat{h})$, and combine the generalized IoU (GIoU) loss $\mathcal{L}_{\text{GIoU}}(b_{pred}, b_{gt})$ (Rezatofighi et al. 2019) and smooth L1 loss $\mathcal{L}_{\text{smooth-L1}}(b_{pred}, b_{gt})$ to optimize the grounding model.

Regional Image Captioning (IC). Similar to the visual grounding, we also formulate the regional image captioning model $\text{Model}_{\text{IC}}(\cdot, \cdot)$ as a black box, which takes image features v and a specific region (denoted by the box coordinate b) as inputs to predict the region-related language expression x_{pred} as follows.

$$x_{pred} = \text{Model}_{\text{IC}}(v, b). \qquad (2)$$

This regional image captioner mainly reuses the image-grounded text decoder in the pre-training stage to generate the language expression. After model pre-training with language modeling (LM) loss, the text decoder already has the zero-shot captioning capability to some extent. Nevertheless, different from the classic image captioner, our model is required to be region-aware. To this end, in the fine-tuning stage, we project the box coordinate b to the regional embedding via a fully-connected layer: $e_{box} = \text{FC}_{\text{box}}(b)$. This regional embedding is added to the vision tokens to obtain the region-aware visual representations $v^\star$:

$$v^\star = v + e_{box}. \qquad (3)$$

Different from the bi-directional self-attention in visual grounding, the captioning model utilizes causal self-attention to facilitate text generation. As shown in Figure 2 (a), a classification head upon Transformer is used

to generate tokens over the vocabulary. For a text sequence $x = \{x_1, x_2, \cdots, x_n\}$, image captioner generates token x_t based on the previous tokens $x_{<t}$ in an auto-regressive fashion. We train the captioning model using unidirectional language modeling objective by maximizing the negative log-likelihood of the text sequence: $-\sum_t \log P_\theta(x_t|v^\star, x_{<t})$, where θ denotes the trainable model parameters.

VG → IC Cycle-Consistency Learning. As shown in Eq. 1 and Eq. 2, visual grounding and regional image captioning perform as the inverse process of each other. Consequently, we can organize them in a cyclic framework for joint optimization using consistency constraints. In the fine-tuning stage, inspired by BLIP (Li et al. 2022), grounding and captioning branches share the model parameters of cross-attention and feed-forward network (FFN) in their Transformer architectures to tightly bridge two individual tasks. This weight-sharing mechanism not only improves training efficiency but also enjoys multi-task learning for mutual prompting.

We first start the cycle-consistency learning from visual grounding (VG) to image captioning (IC), and leverage the training objective of image captioning to optimize two tasks. This process merely needs the language expression x_{init} and visual features v as the inputs, without requiring any bounding box labels as follows.

$$\text{VG:} \quad b_{pred} = \text{Model}_{\text{VG}}(v, x_{init}), \qquad (4)$$

$$\text{IC:} \quad x_{pred} = \text{Model}_{\text{IC}}(v, b_{pred}), \qquad (5)$$

$$\text{Loss:} \quad \mathcal{L}_{\text{VG}\to\text{IC}} = \mathcal{L}_{\text{XE}}(\tilde{x}_{init}, \tilde{x}_{pred}), \qquad (6)$$

where $\mathcal{L}_{\text{XE}}(\cdot, \cdot)$ denotes the cross-entropy loss, $\tilde{x}_{init}$ represents the one-hot vocabulary distribution of the input expression x_{init}, x_{pred} denotes the predicted text sequence and $\tilde{x}_{pred}$ is the corresponding token prediction probability. In this cycle-consistency learning process, we first utilize the grounding model $\text{Model}_{\text{VG}}(\cdot, \cdot)$ to generate the bounding box coordinate b_{pred}, which serves as the pseudo label of captioning model $\text{Model}_{\text{IC}}(\cdot, \cdot)$. Then we can utilize the initial language expression x_{init} as the supervision signal of the generated caption x_{pred} to form the cyclic supervision constraint. In this way, we can optimize the model without providing any bounding box annotations.

IC → VG Cycle-Consistency Learning. We can also build the cycle-consistency learning from image captioning (IC) to visual grounding (VG). This cycle merely requires the visual feature and an initial bounding box b_{init} as the inputs:

$$\text{IC:} \quad x_{pred} = \text{Model}_{\text{IC}}(v, b_{init}), \qquad (7)$$

$$\text{VG:} \quad b_{pred} = \text{Model}_{\text{VG}}(v, x_{pred}), \qquad (8)$$

$$\text{Loss:} \quad \mathcal{L}_{\text{IC}\to\text{VG}} = \mathcal{L}_{\text{GIoU+L1}}(b_{init}, b_{pred}), \qquad (9)$$

where $\mathcal{L}_{\text{GIoU+L1}}(\cdot, \cdot)$ denotes the combination of generalized IoU loss (Rezatofighi et al. 2019) and smooth L1 loss. Ideally, the generated bounding box b_{pred} should be consistent with the initial bounding box b_{init}. To this end, we can optimize the whole network via the visual grounding loss.

Model Training

Both visual grounding and regional image captioning models require the image-text-box triplets for training. Labeling

such well-annotated data is rather time-consuming. However, the aforementioned two learning cycles enable us to optimize the models in a semi-weakly supervised manner, where either the ground-truth language expression or bounding box can be omitted. Besides, based on the same data, we observe that adding cycle-consistency constraints to the classic fully supervised training paradigm can further boost the performance, which shows our cycle-consistency learning framework can better exploit the training data.

Fully Supervised Training. In the supervised training, except for the individual training objectives for grounding and captioning models, we also add the cycle-consistency losses (*i.e.,* $\mathcal{L}_{VG \to IC}$ and $\mathcal{L}_{IC \to VG}$) to regularize the models.

The effectiveness of our framework can be explained in an online data augmentation view. It is well recognized that a picture is worth thousands of words. The proposed cycle-consistency learning framework can be regarded as an incremental training process that iteratively predicts pseudo labels to augment the training data (*e.g.,* diverse referring expressions for a region), and thus further boosts the performance of supervised training.

Semi-weakly Supervised Training. We also validate the potential of our framework by conducting semi-weakly supervised training, where only limited fully-annotated images (with *both* referring expression and bounding box) are available. In the experiments, we empirically set the percentage of fully-annotated data as 20%, and the rest 80% images are annotated with only language expressions *or* only bounding boxes. For fully-annotated images, we compute the standard fully-supervised losses for the 20% image-text-box data. As for the partially labeled data, we compute the IC → VG consistency loss for box-only images, and compute the VG → IC consistency loss for text-only images. All the losses are gathered as the total loss of a batch of training samples. By virtue of VG → IC and IC → VG, the grounding and captioning models mutually annotate the weakly-annotated data for collaborative training. Note that the captioning and grounding models share the same visual encoder (ViT) and most blocks (cross-attention and FFN) of their Transformer blocks. To this end, they both benefit from the VG → IC and IC → VG cycle-consistency training.

Experiments

Datasets and Metrics

Pre-training Data. In the pre-training stage, we collect the image-text pairs from Visual Genome (Krishna et al. 2017), COCO (Lin et al. 2014), SBU (Ordonez, Kulkarni, and Berg 2011), Conceptual 3M (Sharma et al. 2018), and a filtered version of LAION (115M images) (Schuhmann et al. 2021).

RefCOCO, RefCOCO+, and RefCOCOg. We evaluate the visual grounding performance on three prevalent benchmarks including RefCOCO (Yu et al. 2016), RefCOCO+ (Yu et al. 2016), and RefCOCOg (Mao et al. 2016). Following the official setting, RefCOCO and RefCOCO+ are split into the train set, validation set, testA set, and testB set. RefCOCOg includes the train set, validation set, and test set.

We consider a referring expression grounded correctly when its predicted region has at least 0.5 Intersection-over-

Union (IoU) with the ground-truth box. We measure the visual grounding performance in terms of top-1 accuracy.

COCO Caption. We evaluate the image captioning performance on the COCO Karpathy split dataset (Lin et al. 2014; Karpathy and Fei-Fei 2015). To evaluate the performance, we leverage standard metrics in the captioning task including BLEU@4 (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015), and SPICE (Anderson et al. 2016).

Implementation Details

In our framework, the image encoder is initialized from ViT-B/16 pre-trained on the ImageNet (Dosovitskiy et al. 2020), and the text encoders of both visual grounding and image captioning branches are initialized from the official BERT-base (Devlin et al. 2018). In the pre-training stage, the model is trained on 32 V100 GPUs for 20 epochs using a batch size of 2880. We use AdamW optimizer (Loshchilov and Hutter 2017) with a weight decay of 0.05. The learning rate is warmed-up to 3×10^{-4} and decayed linearly with a rate of 0.85. We take random image crops of resolution 224×224 during pre-training.

In the fine-tuning stage, we train the model using a small learning rate of 1×10^{-5} and linearly decay it. The additionally added blocks that are not included in the pre-training stage (*i.e.,* box regression head and box project layer) are randomly initialized. For fair comparisons, following (Deng et al. 2021; Li et al. 2022), the input image resolutions are set to 640×640 and 384×384 when evaluating grounding and captioning tasks, respectively. When combining different datasets, we carefully check the train/test sets to avoid image overlap. The captioning model adopts the beam search strategy (beam size = 3) in all experiments. The proposed cycle-consistency model is fine-tuned for 20 epochs.

In the following experiments, our Cycle-Consistency learning of captioning and grounding framework is denoted as **CyCo**.

Ablation Study

Semi-weakly Supervised Training. Thanks to the cycle-consistency design, our approach is able to train the grounding model using weakly labeled data. As shown in Table 1 (top), the grounding performance is poor when only 20% training data is available. The performance gap between the models trained using 20% data and 100% data is considerable, *e.g.,* 13.4% gap on RefCOCOg test set. In "pseudo label" setting in Table 1, we train a grounding/captioning model using the 20% fully-annotated data and leverage this model to label the rest 80% images. Then, all the 100% data are used to train the model, whose performance can be improved but is still unsatisfactory. Further, by annotating the rest 80% training data in an online manner using our CyCo framework, the performance gap is significantly reduced. Finally, our semi-weakly supervised CyCo model only has a minor gap in comparison to the fully-supervised model.

In Table 1 (bottom), we further ablative the regional image captioning task, which requires the model to describe a specific region. Using only 20% full-annotated data, our

Ablation Study of Visual Grounding						
Training Data		RefCOCO			RefCOCOg	
20% data	80% data	val	testA	testB	val	test
Fully-anno	✗	82.07	87.33	73.20	68.08	66.65
Fully-anno	Pseudo labels	84.13	88.36	76.58	74.64	75.10
Fully-anno	Weakly (CyCo)	86.73	90.27	81.87	76.85	77.81
Fully-anno	Fully-anno	**88.73**	**91.07**	**84.27**	**80.23**	**80.06**

Ablation Study of Regional Image Captioning					
Training Data		RefCOCO testA		RefCOCO testB	
20% data	80% data	M	C	M	C
Fully-anno	✗	29.6	77.5	34.0	131.8
Fully-anno	Pseudo labels	31.3	87.4	35.2	143.5
Fully-anno	Weakly (CyCo)	33.7	98.5	36.6	151.6
Fully-anno	Fully-anno	**34.8**	**104.2**	**37.9**	**156.6**

Table 1: Comparison results of semi-weakly supervised and fully supervised models. Top: comparison of visual grounding performance in terms of top-1 accuracy. Bottom: comparison of regional image captioning performance in METEOR (M) and CIDEr (C).

Supervised	Cycle	RefCOCO		RefCOCO+		RefCOCOg	
		testA	testB	testA	testB	val	test
✓		91.07	84.27	84.40	68.20	80.23	80.06
✓	✓	**91.87**	**85.33**	**87.07**	**69.87**	**81.31**	**81.04**

Table 2: Performance study by adding cycle-consistency learning to the naive supervised VG model. Our "supervised+cycle" utilizes the *same* training splits but adds cycle-consistency losses as the additional regularizations, which steadily improves the results.

semi-weakly supervised captioning model significantly reduces the performance gap in comparison to the supervised counterparts. Our CyCo framework has the potential of absorbing more weakly-labeled data such as object detection images with only box labels or image-text pairs with only caption annotations to further improve the performance.

Fully Supervised Training. We further combine this cycle-consistency learning pipeline with the standard supervised learning. In each training iteration, we not only use the ground-truth labels to supervise the grounding model, but also utilize the VG → IC and IC → VG cycles to regularize the training. As shown in Table 2, based on the *same* training data, adding cycle-consistency learning to the classic supervised learning can further boost performance. This reveals that our framework can better exploit the training data by online data augmentation.

Adding More Training Data For VG. Since the proposed framework jointly optimizes visual grounding and image captioning within a unified framework, we can merge the datasets of both tasks for optimization. As shown in Table 3, adding captioning data further improves the visual grounding results. Note that COCO Karpathy split lacks the region/box annotation, which can be regarded as the weakly-labeled data compared to the image-text-box triplet. These results justify the potential of our framework of absorbing more cheap data to achieve superior performance.

Ground Data	Caption Data	RefCOCO			RefCOCOg	
		val	testA	testB	val	test
✓		89.47	91.87	85.33	81.31	81.04
✓	✓	**89.78**	**92.05**	**85.63**	**82.24**	**82.20**

Table 3: Performance study by adding more training data. Adding weakly labeled images (*e.g.,* COCO-caption without box annotations) improves the grounding performance.

	Training Data	RefCOCOg B@4 / CIDEr	COCO-caption B@4 / CIDEr
BLIP	only COCO	24.1 / 74.6	39.7 / 133.3
BLIP	COCO & RefCOCOg	30.8 / 96.2	38.4 / 130.2
CyCo	COCO & RefCOCOg	**36.6 / 128.5**	**40.6 / 133.9**

Table 4: Performance study on image captioning. Compared to our baseline, CyCo can better leverage the datasets from different domains (COCO-caption and RefCOCOg).

Adding More Training Data For IC. In Table 4, we exhibit the image captioning potentials of our CyCo compared to the baseline BLIP. Without the design of regional embedding, BLIP is not aware of the local region. To this end, combining both COCO-caption and RefCOCOg datasets fails to improve its performance. In contrast, our model can freely switch between the local and global captions (Figure 3) and absorb more (fully- or partially-annotated) data to further boost the performance.

State-of-the-art Comparison

Evaluation on RefCOCO/RefCOCO+/RefCOCOg. Table 5 reports the comparison results of state-of-the-art methods on the RefCOCO, RefCOCO+, and RefCOCOg datasets. As ablated in Table 3, adding more captioning data can steadily boost the grounding performance. For fairness, when fine-tuning the visual grounding model, we only use the standard training splits of the above benchmarks without involving other training data. Thanks to the vision-language pre-training, our approach significantly outperforms the classic grounding methods without pre-training. We also include some representative pre-training-based approaches including UNITER, VILLA, ERNIE-ViL, and BLIP to justify the superior performance of our approach. For example, our method surpasses VILLA and ERNIE-ViL on the RefCOCO and RefCOCO+. On the RefCOCOg dataset with longer expressions, our method also exhibits satisfactory results. In Table 5, we also present some high-performance methods (*e.g.,* MDETR, OFA, and FIBER) that additionally leverage the expensive image-text-box data for pre-training. Besides, FIBER adopts a strong visual backbone and higher image resolution (*e.g.,* $800 \times 1,333$), exhibiting the leading performance. We can also improve our performance by adopting a stronger backbone and higher image resolution, which leave as our future work.

Furthermore, our framework enables model training using partially labeled data. In the semi-weakly supervised setting (*e.g.,* 20% fully-annotated data and 80% weakly-annotated data), it is worth noting that our method also steadily out-

Method	Backbone	Supervised	Data		RefCOCO			RefCOCO+			RefCOCOg	
			I-T	I-T-B	val	testA	testB	val	testA	testB	val	test
w/o Pre-training												
MCN (Luo et al. 2020)	DarkNet-53	✓			80.08	82.29	74.98	67.16	72.86	57.31	66.46	66.01
QRNet (Ye et al. 2022)	Swin-S	✓			84.01	85.85	82.34	72.94	76.17	63.81	73.03	72.52
VLTVG (Yang et al. 2022)	ResNet-101	✓			84.77	87.24	80.49	74.19	78.93	65.17	76.04	74.18
VG-LAW (Su et al. 2023)	ViT-B	✓			86.62	89.32	83.16	76.37	81.04	67.50	76.90	76.96
TransVG++ (Deng et al. 2023)	ViT-B	✓			86.28	88.37	80.97	75.39	80.45	66.28	76.18	76.30
Img-Text Pre-training												
UNITER-B (Chen et al. 2020b)	ResNet-101	✓	✓		81.24	86.48	73.94	75.31	81.30	65.58	74.31	74.51
VILLA-B (Gan et al. 2020)	ResNet-101	✓	✓		81.65	87.40	74.48	76.05	81.65	65.70	75.90	75.93
ERNIE-ViL-B (Yu et al. 2020)	ResNet-101	✓	✓		-	-	-	74.02	80.33	67.74	-	-
BLIP-B⋆ (Li et al. 2022)	ViT-B	✓	✓		88.73	91.07	84.27	79.53	84.40	68.20	80.23	80.06
Img-Text-Box Pre-training												
MDETR (Kamath et al. 2021)	ResNet-101	✓	✓	✓	86.75	89.58	81.41	79.52	84.09	70.62	81.64	83.31
OFA-L (Wang et al. 2022)	BART-L	✓	✓	✓	90.05	92.93	85.26	84.49	90.10	77.77	84.54	85.20
FIBER-B (Dou et al. 2022)	Swin-B	✓	✓	✓	90.68	92.59	87.26	85.74	90.13	79.38	87.11	87.32
CyCo (semi-weakly)	ViT-B	✗	✓		86.73	90.27	81.87	72.53	82.00	61.33	76.85	77.81
CyCo (fully supervised)	ViT-B	✓	✓		**89.47**	**91.87**	**85.33**	**80.40**	**87.07**	**69.87**	**81.31**	**81.04**

Table 5: Comparisons with state-of-the-art methods on RefCOCO, RefCOCO+, and RefCOCOg in terms of top-1 accuracy (%). The BLIP-B⋆ denotes our implemented BLIP-B model on the visual grounding task, which is not included in the original BLIP framework (Li et al. 2022). In the pre-training stage, we only use the image-text (I-T) pairs without any image-text-box (I-T-B) triplets. We report the performance of our semi-weakly supervised model as well as the supervised model.

Method	Cross-Entropy			
	B@4	M	C	S
w/o Pre-training				
AoANet (Huang et al. 2019)	37.2	28.4	119.8	21.3
X-LAN (Pan et al. 2020)	38.2	28.8	122.0	21.9
w/ Pre-training				
Oscar-B (Li et al. 2020)	36.5	30.3	123.7	23.1
VinVL-B (Zhang et al. 2021)	38.2	30.3	129.3	23.6
LEMON-B (Hu et al. 2021a)	40.3	30.2	133.3	23.3
BLIP-B (Li et al. 2022)	39.7	-	133.3	23.3
SimVLM-B (Wang et al. 2021b)	39.0	**32.9**	**134.8**	24.0
CyCo (Ours)	**40.6**	31.2	133.9	**24.4**

Table 6: Performance comparisons on the COCO-caption Karpathy test split, where B@4, M, C, S denote BLEU@4, METEOR, CIDEr, and SPICE scores, respectively.

performs the previous fully supervised counterparts such as UNITER and VILLA on the RefCOCO and RefCOCOg.

Evaluation on COCO Caption. Mainstream image captioning methods focus on describing the global image context. However, optimizing the proposed framework only on the grounding dataset will overlook the captioning capability of the global content. To this end, we merge two datasets including COCO Karpathy train split and RefCOCOg train split to jointly optimize our cycle-consistency framework. We define the referred region of COCO-caption dataset as the whole image. After joint training, we assess the global captioning performance on the COCO Karpathy test split.

In Table 6, compared with the state-of-the-art captioning methods with pre-training such as Oscar and VinVL, our method shows better results. Our method is even comparable with the recent SimVLM trained with 1.8 billion image-text pairs, which is 10× larger than ours. The proposed method
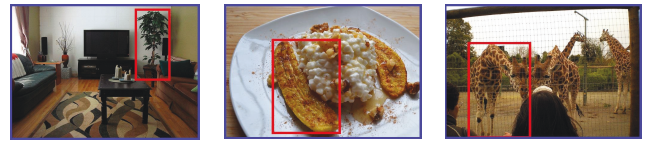
Figure 3: Our captioner can describe both the local (red region) and global (blue region) contexts.

slightly outperforms our baseline BLIP.

It is worth noting that our captioner focuses on a more challenging scenario and can describe arbitrary image regions, which is infeasible for all the comparison methods. Figure 3 showcases some examples. When describing a specific region, our captioner captures its relationship with surrounding objects to avoid ambiguous expressions. When describing the global image content, the proposed method performs in a similar way as the general image captioners.

Conclusion

In this paper, we bridge two vision-language tasks including visual grounding and regional image captioning in a cyclic training pipeline. By cycle-consistency constraints, the proposed framework can exploit the weakly-annotated data for model training and augment the training samples to further boost the fully supervised training. Extensive experiments on image captioning and visual grounding datasets verify the effectiveness of our framework.

References

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. Spice: Semantic propositional image caption evaluation. In *ECCV*.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL Workshop*.

Chen, S.; Jin, Q.; Wang, P.; and Wu, Q. 2020a. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In *CVPR*.

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020b. Uniter: Universal image-text representation learning. In *ECCV*.

Cornia, M.; Baraldi, L.; and Cucchiara, R. 2019. Show, control and tell: A framework for generating controllable and grounded captions. In *CVPR*.

Deng, J.; Yang, Z.; Chen, T.; Zhou, W.; and Li, H. 2021. Transvg: End-to-end visual grounding with transformers. In *ICCV*.

Deng, J.; Yang, Z.; Liu, D.; Chen, T.; Zhou, W.; Zhang, Y.; Li, H.; and Ouyang, W. 2023. Transvg++: End-to-end visual grounding with language conditioned vision transformer. *IEEE TPAMI*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Dou, Z.-Y.; Kamath, A.; Gan, Z.; Zhang, P.; Wang, J.; Li, L.; Liu, Z.; Liu, C.; LeCun, Y.; Peng, N.; et al. 2022. Coarse-to-Fine Vision-Language Pre-training with Fusion in the Backbone. *arXiv preprint arXiv:2206.07643*.

Gan, Z.; Chen, Y.-C.; Li, L.; Zhu, C.; Cheng, Y.; and Liu, J. 2020. Large-scale adversarial training for vision-and-language representation learning. In *NeurIPS*.

Guo, L.; Liu, J.; Yao, P.; Li, J.; and Lu, H. 2019. Mscap: Multi-style image captioning with unpaired stylized text. In *CVPR*.

He, D.; Xia, Y.; Qin, T.; Wang, L.; Yu, N.; Liu, T.-Y.; and Ma, W.-Y. 2016. Dual learning for machine translation. In *NeurIPS*.

Hu, X.; Gan, Z.; Wang, J.; Yang, Z.; Liu, Z.; Lu, Y.; and Wang, L. 2021a. Scaling up vision-language pre-training for image captioning. *arXiv preprint arXiv:2111.12233*.

Hu, X.; Yin, X.; Lin, K.; Wang, L.; Zhang, L.; Gao, J.; and Liu, Z. 2021b. Vivo: Surpassing human performance in novel object captioning with visual vocabulary pre-training. In *AAAI*.

Huang, L.; Wang, W.; Chen, J.; and Wei, X.-Y. 2019. Attention on attention for image captioning. In *ICCV*.

Johnson, J.; Karpathy, A.; and Fei-Fei, L. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*.

Kamath, A.; Singh, M.; LeCun, Y.; Synnaeve, G.; Misra, I.; and Carion, N. 2021. MDETR-modulated detection for end-to-end multi-modal understanding. In *ICCV*.

Karpathy, A.; and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

Kazemzadeh, S.; Ordonez, V.; Matten, M.; and Berg, T. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*.

Kinghorn, P.; Zhang, L.; and Shao, L. 2018. A region-based image caption generator with refined descriptions. *Neurocomputing*, 272: 416–424.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1): 32–73.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. *arXiv preprint arXiv:2201.12086*.

Li, W.; Gao, C.; Niu, G.; Xiao, X.; Liu, H.; Liu, J.; Wu, H.; and Wang, H. 2021. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. In *ACL*.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Liu, D.; Zhang, H.; Wu, F.; and Zha, Z.-J. 2019. Learning to assemble neural module tree networks for visual grounding. In *ICCV*.

Liu, Y.; Wan, B.; Ma, L.; and He, X. 2021. Relation-aware instance refinement for weakly supervised visual grounding. In *CVPR*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Luo, G.; Zhou, Y.; Sun, X.; Cao, L.; Wu, C.; Deng, C.; and Ji, R. 2020. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *CVPR*.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A. L.; and Murphy, K. 2016. Generation and comprehension of unambiguous object descriptions. In *CVPR*.

Ordonez, V.; Kulkarni, G.; and Berg, T. 2011. Im2text: Describing images using 1 million captioned photographs. In *NeurIPS*.

Pan, Y.; Yao, T.; Li, Y.; and Mei, T. 2020. X-linear attention networks for image captioning. In *CVPR*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; and Savarese, S. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*.

Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; and Komatsuzaki, A. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Shah, M.; Chen, X.; Rohrbach, M.; and Parikh, D. 2019. Cycle-consistency for robust visual question answering. In *CVPR*.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*.

Su, W.; Miao, P.; Dou, H.; Wang, G.; Qiao, L.; Li, Z.; and Li, X. 2023. Language adaptive weight generation for multi-task visual grounding. In *CVPR*.

Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; and Dai, J. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. In *ICLR*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *CVPR*.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *CVPR*.

Wang, L.; Huang, J.; Li, Y.; Xu, K.; Yang, Z.; and Yu, D. 2021a. Improving weakly supervised visual grounding by contrastive knowledge distillation. In *CVPR*.

Wang, N.; Song, Y.; Ma, C.; Zhou, W.; Liu, W.; and Li, H. 2019. Unsupervised deep tracking. In *CVPR*.

Wang, N.; Xie, J.; Luo, H.; Cheng, Q.; Wu, J.; Jia, M.; and Li, L. 2023a. Efficient image captioning for edge devices. In *AAAI*.

Wang, N.; Xie, J.; Wu, J.; Jia, M.; and Li, L. 2023b. Controllable image captioning via prompting. In *AAAI*.

Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai, S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; and Yang, H. 2022. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*.

Wang, X.; Jabri, A.; and Efros, A. A. 2019. Learning correspondence from the cycle-consistency of time. In *CVPR*.

Wang, Y.; Xu, J.; and Sun, Y. 2022. End-to-End Transformer Based Model for Image Captioning. In *AAAI*.

Wang, Z.; Yu, J.; Yu, A. W.; Dai, Z.; Tsvetkov, Y.; and Cao, Y. 2021b. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.

Yang, L.; Xu, Y.; Yuan, C.; Liu, W.; Li, B.; and Hu, W. 2022. Improving Visual Grounding with Visual-Linguistic Verification and Iterative Reasoning. In *CVPR*.

Yang, S.; Li, G.; and Yu, Y. 2019. Dynamic graph attention for referring expression comprehension. In *ICCV*.

Yang, Z.; Gong, B.; Wang, L.; Huang, W.; Yu, D.; and Luo, J. 2019. A fast and accurate one-stage approach to visual grounding. In *ICCV*.

Ye, J.; Tian, J.; Yan, M.; Yang, X.; Wang, X.; Zhang, J.; He, L.; and Lin, X. 2022. Shifting More Attention to Visual Backbone: Query-modulated Refinement Networks for End-to-End Visual Grounding. In *CVPR*.

Yu, F.; Tang, J.; Yin, W.; Sun, Y.; Tian, H.; Wu, H.; and Wang, H. 2020. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *arXiv preprint arXiv:2006.16934*.

Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling context in referring expressions. In *ECCV*.

Yu, L.; Tan, H.; Bansal, M.; and Berg, T. L. 2017. A joint speaker-listener-reinforcer model for referring expressions. In *CVPR*.

Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; and Gao, J. 2021. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*.

Zhou, L.; Palangi, H.; Zhang, L.; Hu, H.; Corso, J.; and Gao, J. 2020. Unified vision-language pre-training for image captioning and vqa. In *AAAI*.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*.