FindIt: Generalized Localization with Natural Language Queries

Weicheng Kuo, Fred Bertsch, Wei Li, AJ Piergiovanni, Mohammad Saffar, Anelia Angelova

Google Research, Brain Team
{weicheng,fredbertsch,mweili,ajpiergi,msaffar,anelia}@google.com

Abstract. We propose FindIt, a simple and versatile framework that unifies a variety of visual grounding and localization tasks including referring expression comprehension, text-based localization, and object detection. Key to our architecture is an efficient multi-scale fusion module that unifies the disparate localization requirements across the tasks. In addition, we discover that a standard object detector is surprisingly effective in unifying these tasks without a need for task-specific design, losses, or pre-computed detections. Our end-to-end trainable framework responds flexibly and accurately to a wide range of referring expression, localization or detection queries for zero, one, or multiple objects. Jointly trained on these tasks, FindIt outperforms the state of the art on both referring expression and text-based localization, and shows competitive performance on object detection. Finally, FindIt generalizes better to out-of-distribution data and novel categories compared to strong single-task baselines. All of these are accomplished by a single, unified and efficient model. The code will be released.[1]

1 Introduction

Natural language enables flexible descriptive queries about images. The interaction between text queries and images grounds linguistic meaning in the visual world, facilitating a stronger understanding of object relationships, human intentions towards objects, and interactions with the environment. The research community has studied visual grounding through tasks including phase grounding, object retrieval and localization, language-driven instance segmentation, and others [62,70,60,68,56,80,25,21].

Among the most popular visual grounding tasks is referring expression comprehension (REC), which localizes an object given a referring text [90,55,70]. This task often requires complex reasoning on prominent objects. A highly related semantic localization task is object detection (DET), which seeks to detect all objects from a predefined set of classes without text inputs [58,69,75,78,17,66]. In contrast to REC, this task requires the accurate classification and localization

[1] Please see the project page: https://sites.google.com/view/findit-eccv22/home.

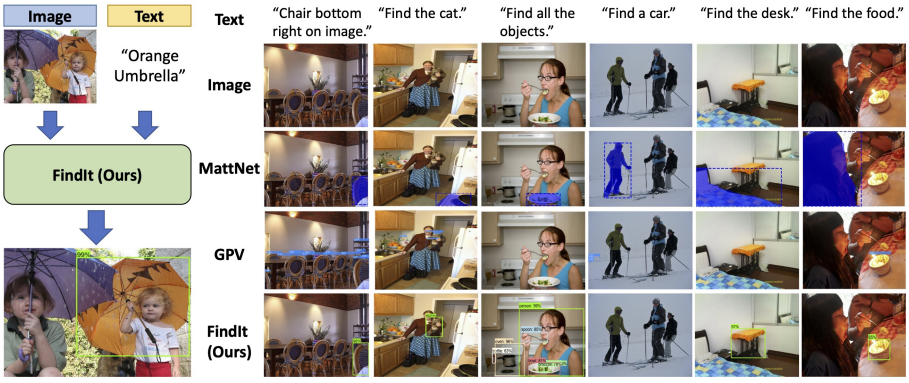

Fig. 1: FindIt is a general-purpose model for visual grounding and localization tasks (left). The input is an image-text pair specifying the objects of interest using natural language, and the outputs are a set of bounding boxes and classification scores. Specifically, FindIt addresses the following tasks (col. 1-3): referring expression comprehension (col. 1), text-based localization (col. 2), and the object detection task by an optional generic prompt e.g. "Find all the objects.", (col. 3). Furthermore, FindIt can respond accurately when the referred object is absent (col. 4), or when it is tested on out-of-distribution (OOD) images and with novel category names, e.g. "desk", where "dining table" is the closest category in the training set (col. 5). FindIt can also locate objects, referred to by novel super-category names e.g. "food" (col. 6). We compare to MattNet [89] and GPV [20] in all these scenarios. (Best viewed in color)

of small, occluded objects. At the intersection of the two is text-based localization [20,23] (LOC), in which a simple category-based text query prompts the model to detect the objects of interest.

Due to their highly dissimilar task properties, REC, DET, and LOC are mostly studied through separate benchmarks with most models only dedicated to one task [87,67,20]. As a result, existing models have not adequately synthesized information from the three tasks to achieve a more holistic visual and linguistic understanding. REC models, for instance, are trained to predict one object per image, and often struggle to localize multiple objects[2], reject negative queries, or detect novel categories (see Figure 1). In addition, DET models are unable to process text inputs, and LOC models often struggle to process complex queries such as "Chair bottom right on image" (see Figure 1). Lastly, none of the models can generalize sufficiently well beyond the their training data and categories.

To address these limitations, we propose a unified visual localization approach which we call FindIt. Key to our architecture is a multi-level cross-modality fusion module which can perform complex reasoning for REC and simultaneously

[2] Technically, many REC models can localize more than one object, but they often struggle because they are only trained to predict one object per image on REC data.

recognize small and challenging objects for LOC and DET. To unify the disparate demands of these tasks, the module efficiently fuses and learns features across many levels of abstraction. Concretely, we utilize the more expressive cross-attention fusion on lower resolution features, and the more efficient product fusion on higher resolution features to combine the best of both worlds. Last but not least, we discover that a standard object detector and detection losses [67] are sufficient and surprisingly effective for REC, LOC, and DET tasks without a need for task-specific design and losses [12,20,50,54,86,87,89]. In short, FindIt is a simple, efficient, and end-to-end trainable model for unified visual grounding and object detection.

By learning REC, LOC, and DET jointly in one model, FindIt acquires a more holistic and versatile capability for visual grounding than its single-task counterparts. Notably, FindIt surpasses the state of the art on REC and LOC, and demonstrates competitive performance on DET. Moreover, unlike existing task-specific models, FindIt accomplishes these in a single model that can respond flexibly to a wide range of referring expression and localization queries, solve the standard detection task, and generalize better to novel data and classes. In summary, our contributions are:

- We propose FindIt, a simple and versatile framework for visual grounding and detection tasks. In contrast to task-specific models, a single FindIt model can respond flexibly to a wide range of referring expression and localization queries, solve the standard detection task, and generalize better to novel data and classes.
- We propose an efficient multi-scale cross-attention fusion module to unify the disparate task requirements between REC, LOC, and DET. Using the fused features, we discover that a standard detector and detection losses are surprisingly effective for all tasks without a need for task-specific design or losses.
- We surpass the state of the art on REC and LOC, and show competitive DET performance within a single, unified and efficient model.

2 Related Work

Referring Expression Comprehension (REC) and phrase grounding tasks [55,76,90,89,63,87,35,86,62] require the models to ground linguistic elements in the image. Several datasets which enable and enrich the study of these tasks have been proposed [90,55,30,33,70,9,62]. Yu et al. and Mao et al. [90,55] expand the COCO benchmark with referring expression annotations, while the Referit game [70] crowd-sources such labels through game-play. One-stage [7,43,87,12,54] and two-stage [92,89,26,81,85,49] methods have been popular for these tasks. **Object Detection (DET)** task is well established and has a plethora of approaches [17,66,67,22,45] and benchmarks [46,71]. The goal is to identify the bounding boxes of a set of pre-defined classes without prompting by text. Many recent approaches have started to study the open-set and zero-shot settings [14,2,82,96,19]. **Text-based Localization (LOC)** has been recently proposed

alongside other vision and language tasks [20,23]. Text-based localization is similar to the referring expression comprehension task. The text query specifies an object class to be localized. This task is typically derived from standard detection datasets [46,38]. Early results with this tasks are presented by [20,23] where the focus has been on a single object [23]. FindIt extends this capability to localize multiple objects of any given category or detect all objects of a given vocabulary through a free-form text prompt.

Multi-modal Vision-Language Learning. Large amounts of vision and language work are present, such as visual-grounding [15,79,48,94,89,60], image captioning [6,1,5], visual question and answering (VQA) [32,11], visual reasoning [73,91,83], image-text retrieval [31,64,72], and video-text learning [28,47,34,11]. Many approaches to vision-language learning leverage large-scale image-text pretraining or pre-computed detections [52,40,74,8,42,29,84,95,37,88,13,51,5,64]. In particular, many methods underscore the importance of localization to increase the success of related vision-and-language understanding/reasoning tasks such as VQA and CLEVR [8,42,93,16,1,4,35].

Vision and Language Feature Fusion. Recently, the Transformer [77] and its cross-modality variants [52,8,36] have been popular fusion choices for vision-language tasks. To localize objects at various scales, existing REC works have used multi-level fusion by applying activation and product fusion [54,86] or concatenation and convolution fusion [87]. Inspired by recent works [12,8,52,36] on single-scale cross-attention, we propose multi-scale fusion to satisfy the disparate requirements of REC and detection tasks, where REC requires complex reasoning while detection requires accurate localization and recognition. The fusion module enables us to unify these tasks in a single model and surpass the state of the art on REC, LOC and maintains competitive DET performance.

Multi-task Learning for Visual Grounding and Object Localization. Existing approaches have combined grounding and localization tasks with text-generation tasks such as VQA, captioning, visual entailment [53,10,27], and have leveraged pretraining or joint training with similar localization tasks [54,41,35]. Hu et al. [27] unifies a detection task with text-generation tasks. GPV [20] combines text-based localization with VQA by generating both boxes and text for each input image/text pairs. MCN [54] jointly learn REC and RES (Referring Expression Segmentation) to show the benefits of multi-task learning for both. GLIP [41] formulates object detection as phrase grounding and combines detection, caption, and grounding datasets for zero/few-shot detection. M-DETR [35] uses many grounding datasets in a phrase grounding pretraining. Similar to MCN, FindIt unifies semantically similar tasks to study the benefits of multi-task learning. Different from M-DETR and GLIP, FindIt uses only COCO and RefCOCO data without a need for pretraining on external data.

3 Method

3.1 Overview

The goal of FindIt is to unify a family of semantically-related localization tasks: 1) referring expression comprehension (REC), 2) text-based localization (LOC),

Table 1: FindIt tasks comparison. FindIt unifies the referring expression (REC), text-based localization (LOC) and detection (DET) tasks.

Task	Text Input	Output	Image Size	Loss and Architecture	Metric
REC	Expr. for one object	One box	256 [87,59] / 640[12]	Ref-specific or DETR loss/arch. [87,59,12]	Precision
LOC	Expr. for one class	Many boxes	640 [20]	DETR loss and DETR + image-text fusion [20]	AP50
DET	None / Task prompt	Many boxes/classes	1333 [22,18]	Two-stage [67], one-stage [66,45], transformer [3]	AP
FindIt	All the above	Many boxes/classes	640	Two-stage detector loss [67] + image-text fusion	All

and 3) detection (DET). To accomplish this, FindIt produces a set of boxes/classes when given an RGB image and a text query (see Table 1). The architecture (Section 3.3) includes an image encoder, a text encoder, a fusion model, and a set of box/class prediction heads. The fusion model (Section 3.4) takes multi-scale features from the image encoder and fuses them with the text encoder features. The box/class heads take the fused features as input and produce a set of bounding boxes, their categories and confidence scores. All tasks share the same architecture, losses, and weights.

3.2 Task Definitions

Table 1 shows a comparison of the FindIt sub-tasks. Since these tasks are similar in nature, our goal is to unify, and consider them jointly. We define them as follows:

- **REC**: In the referring expression comprehension task, inputs are an image and a user query about a specific (often prominent) object in the image. The expected output is one bounding box around the correct object. While natural queries may invoke multiple objects, this task is limited to providing a single box as an answer. We adopt the standard precision@1 metric.
- **LOC**: In the text-based localization task, inputs are an image and a query about a category, e.g. "Find the cars" [20]. The expected output is a set of bounding boxes around all objects in that category. This task challenges the model to only predict the relevant objects based on the query. We follow the AP50 metric proposed by [20].
- **DET**: In the detection task, inputs are an image and a standard query, "Find all the objects". The expected outputs are bounding boxes around the objects of categories present in the dataset and their classes, but as we show in Table 4, FindIt can generalize to novel categories via text-based localization. Our modification allows us to share the same vision and language interface with the other tasks. We adopt the standard mAP metric in detection [46].

3.3 Network Architecture

Our network architecture is simple and extensible: it includes an image encoder, a text encoder, a fusion model, and box/class prediction heads (Figure 2). All parameters are shared by all tasks, i.e. there are *no task-specific parameters*. The image encoder is a ResNet backbone which yields multi-level features. The

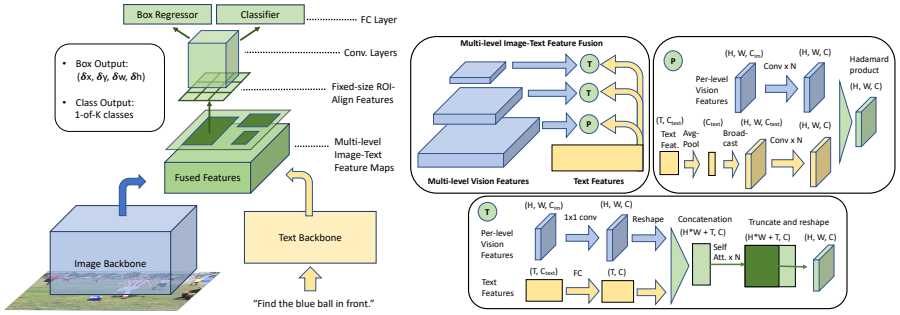

Fig. 2: (Left) Our main architecture accepts an image and a query text as inputs, and processes them separately in image/text backbones before applying the multi-level fusion. We feed the fused features to region proposal network to generate candidate regions and then extract the region features for box regression and classification. (Right) Our multi-level image-text fusion module (top-left) uses transformer fusion blocks (T), and product fusion blocks (P) at the higher/lower levels of the feature maps respectively.

text encoder is a T5 transformer [65] model which encodes a query sentence as a series of token features. The fusion model fuses the multi-level image features with token features (Section 3.4). We fuse the image and text features at the image level, as it allows more flexibility to adapt visual representation to various queries. After the fusion, we apply the standard region proposer [67] and box/class decoders [67]. Our design can tackle any task that predicts multiple objects and their classes given an image and a text query (optional). Although we use FRCNN [67] in this work, our unification approach is agnostic to the choice of detectors and other detectors are also viable [3,45,66][3].

3.4 Multi-level Image-Text Fusion

To combine these different localization tasks, one major challenge is that they are created around different domains and with different goals (see Table 1). For example, the referring expression task primarily references prominent objects in the image rather than small, occluded or faraway objects such that low resolution images would suffice. In contrast, the detection task aims to detect objects with various sizes and occlusion levels in higher resolution images. Apart from these benchmarks, the general visual grounding problem is inherently multiscale, as natural queries can refer to objects of any size. This motivates our multi-level image-text fusion model for efficient processing of higher resolution images over different localization tasks.

We fuse multi-level image features with the text features using a Transformer-based cross attention module [77] (See Figure 2). The vision features at each level are fused with the text features. A feature pyramid [44] fuses features across

[3] The detector head may also be adapted from existing visual grounding models such as [12,87], but we leave this for future studies.

resolutions by progressively up-sampling the higher level fused features to the resolution of lower level features.

The transformer fusion works as follows (see bottom right of Figure 2). We first use a linear layer to project the vision and text features into the same dimension at each level. Next, we collapse the spatial dimension of vision features into a sequence and concatenate it with the text sequence features. We compute the relative position bias based on the total length of the concatenated sequence before applying the self-attention layers. As self-attention is intractable with large feature maps, we apply product fusion (see top right of Figure 2) for the early high resolution feature maps (i.e. F2 and F3), and use self-attention for the smaller, higher level feature maps (i.e. F4 and F5). Ablation studies show the benefits of multi-level fusion and self-attention for handling complex queries (see Section 4.3). Finally, we truncate and reshape the fused features to the same spatial dimensions as the input vision features.

3.5 Task Unification and Multi-task Learning

The three localization tasks must be unified in terms of model, loss, and inputs so they can be trained together. The implications of unification are significant. First, all tasks can share the same model during both training and inference time. Second, the unification of inputs and loss enables us to efficiently train on multiple datasets. Lastly, the model can leverage information from other tasks, which allows the transfer of visual concepts and enables zero-shot applications. For example, we can learn long-tail concepts from the referring expression task and transfer them to other localization tasks.

Apart from the unified architecture (see Section 3.3 and 3.4), datasets are adapted to the different tasks as follows. For the localization task, detection datasets are adapted by generating a set of queries over the categories present in the image. For any present category, the text query takes the form "Find the X" where X is the category name. The objects corresponding to that category are labeled as foreground and the other objects as background. At training time, we randomly sample a text query and corresponding objects from each image. For the detection task, detection datasets are adapted by adding a static task prompt such as "Find all the objects". We found that the specific choice of prompts are not important for LOC and DET tasks (see Table 6a).

After adaptation, all tasks in consideration share the same inputs and outputs—an image input, a text query, and a set of output bounding boxes/classes. We then combine the datasets and train on the mixture. At training time, we use a mixing ratio of 1:1:1 between DET:LOC:REC tasks in each minibatch. To ensure each dataset is sampled adequately, we use a larger batch size of 256 split among the 3 tasks. To make the image size uniform across tasks (see Table 1), we adopt the LOC task's image size of 640 [20] as a middle ground. This is larger but comparable to the image size of REC task [87,86,12]. It is smaller than the size of DET task's images [22] which might limit performance on smaller objects.

Finally, we unify the losses of all tasks. The losses we use are box classification and regression loss, region proposal classification and regression loss, and

weight decay. The loss formulation and relative weights follow [67] without any task-specific modification. All losses have equal weights across tasks. We note that it is unclear how to use existing grounding models out-of-the-box for task unification due to the task-specific architectures, losses, and training strategies [12,20,50,54,86,87,89].

3.6 Implementation details

FindIt uses a region proposer (RPN) [67], class predictor [67] and a class-agnostic box regressor [67] shared among all tasks. The class decoder has the same number of outputs as the detection vocabulary size (i.e. 80 for COCO), as it primarily serves the detection task. We note that no pre-computed detections are used in FindIt as in many two-stage referring expression models [89,50].

FindIt image encoder is initialized from the ResNet50 model pretrained on COCO detection. FindIt text encoder is initialized from T5-base [65] pretrained checkpoint. All other modules are trained from scratch, including the multi-level fusion model, feature pyramid network [44], the region proposal network (RPN) and the box/class decoders [67]. All hyper-parameters of the feature pyramid, RPN and box/class decoder heads follow the Faster R-CNN [18].

We set the batch size to 256 split among 3 tasks DET:LOC:REC with mixing ratio 1:1:1 in the minibatch. The ratio was chosen for simplicity and has room for further optimization. We train the model for 150k steps on a learning rate of 0.08, linear warmup of 500 steps, and a decay factor of 0.1 at 70% and 90% of the training schedule. Total training takes about 1.2 days. For the ablations, we train for 25k steps (0.25x) on the same learning rate schedule. We set the learning rate of the pretrained image encoder and text encoder to be 10% of the rest of the model which trains from scratch [86,12].

We apply random scale jittering uniformly sampled between [0.4, 2.5] for every input image. The image is padded or randomly cropped to the size of (640, 640) after the scale jittering. For the ablation studies, we reduce the scale jittering magnitude to [0.8, 1.25] due to the shorter training. For detection and text-based localization tasks we also apply random horizontal flip following the standard protocol [67]. In addition, we tokenize the text with SentencePiece [39] following T5 [65] and set the maximum expression length to 64 for all tasks.

4 Experiments

We compare FindIt to the state of the art (SOTA) on REC, LOC and DET tasks (Section 4.1). We follow the protocols established in prior works [20,90], using only MS-COCO [46] for training and validation. In addition, we evaluate how FindIt generalizes to OOD datasets and settings (Section 4.2).

Here we define the family of FindIt models. *FindIt* is trained jointly on REC, LOC, and DET tasks, while *FindIt-REC, FindIt-LOC, FindIt-DET* are trained on each individual task to serve as single-task baselines. FindIt does not require more labeled data than existing REC methods, because pre-trained detector outputs [89,92] or initialization with detector weights [87,86,12] have

Table 2: Comparison with state-of-the-art methods on RefCOCO including those using external data for pretraining. We outperform the state of the art on RefCOCO [90], RefCOCO+ [90] and RefCOCOg [55] with only R50 backbone. FindIt-REC is our own single task baseline. **Bold** indicates the highest non-unified training number. ***Red*** indicates the highest number overall, whereas ***blue*** the second highest. (Best viewed in color)

Models	Backbone	RefCOCO			RefCOCO+			RefCOCOg		
		val	testA	testB	val	testA	testB	val-g	val-u	test-u
Two-stage:										
CMN [26]	VGG16	-	71.03	65.77	-	54.32	47.76	57.47	-	-
VC [92]	VGG16	-	73.33	67.44	-	58.40	53.18	62.30	-	-
ParalAttn [97]	VGG16	-	75.31	65.52	-	61.34	50.86	58.03	-	-
MAttNet [89]	R101	76.65	81.14	69.99	65.33	71.62	56.02	-	66.58	67.27
LGRANs [81]	VGG16	-	76.60	66.40	-	64.00	53.40	61.78	-	-
DGA [85]	VGG16	-	78.42	65.53	-	69.07	51.99	-	-	63.28
RvG-Tree [24]	R101	75.06	78.61	69.85	63.51	67.45	56.66	-	66.95	66.51
NMTree [49]	R101	76.41	81.21	70.09	66.46	72.02	57.52	64.62	65.87	66.44
CM-Att-Erase [50]	R101	78.35	83.14	71.32	68.09	73.65	58.03	-	67.99	68.67
One-stage:										
SSG [7]	DarkNet-53	-	76.51	67.50	-	62.14	49.27	47.47	58.80	-
FAOA [87]	DarkNet-53	72.54	74.35	68.50	56.81	60.23	49.60	56.12	61.33	60.36
RCCF [43]	DLA-34	-	81.06	71.85	-	70.35	56.32	-	-	65.73
ReSC-Large [86]	DarkNet-53	77.63	80.45	72.30	63.59	68.36	56.81	63.12	67.30	67.20
MCN [54]	DarkNet-53	80.08	82.29	74.98	67.16	72.86	57.31	-	66.46	66.01
Transformer:										
TransVG [12]	R50	80.32	82.67	78.12	63.50	68.15	55.63	66.56	67.66	67.44
TransVG [12]	R101	81.02	82.72	78.35	64.82	70.70	56.94	67.02	68.67	67.73
FindIt-REC	R50	79.45	82.43	72.94	66.01	71.13	58.62	63.91	67.73	68.77
FindIt	R50	**84.66**	**85.50**	**83.46**	**73.85**	**78.57**	**67.31**	**73.25**	**77.64**	**77.02**
Unified Training:										
FindIt-MIX	R50	84.92	85.54	83.44	74.31	76.93	69.91	82.77	83.17	84.11
FindIt-MIX (384)	R50	**87.09**	85.55	**86.89**	76.35	75.47	**71.85**	**89.84**	**90.40**	**91.01**
FindIt-MIX (384)	R101	**87.91**	86.56	**88.04**	**77.24**	77.42	**73.12**	**90.58**	**90.97**	**91.72**
External Data:										
UNITER-L [8]	R101	81.41	87.04	74.17	75.90	81.45	66.70	-	74.86	75.77
VILLA-L [16]	R101	82.39	**87.48**	74.84	76.17	**81.54**	66.84	-	76.18	76.71
MDETR [35]	R101	86.75	**89.58**	81.41	**79.52**	**84.09**	70.62	-	81.64	80.89

been commonly used. Towards further unification, *FindIt-MIX* trains on all RefCOCO splits (as opposed to a single RefCOCO split used by *FindIt*), LOC, and DET together, resulting in one model for all splits instead of one model for each split, which is the case with FindIt. To our best knowledge, we are the first to report single-model unified training results on RefCOCO benchmarks. We report all FindIt-MIX (384) results as an average of five independent runs.

4.1 Main results

Table 2, Table 3a, and Table 3b show our results on REC, LOC and DET tasks compared to the SOTA. In each table, we compare FindIt to both single- or multi-task approaches for the corresponding task. The single-task approaches are advantaged as they are fully optimized for the task.

Table 2 compares with existing COCO-trained methods on the three popular REC benchmarks: RefCOCO [90], RefCOCO+ [90] and RefCOCOg [55]. We see that FindIt outperforms the SOTA results, including two-stage/one-stage methods and recent Transformer-based models. In particular, on the challenging splits of RefCOCO+ (no location-based information) and RefCOCOg (longer expressions), FindIt outperforms the SOTA results by a clear margin of 5-10

Table 3: Text-based Localization and Detection Benchmarks. All models in the tables use the ResNet-50 (R50) backbone.

(a) Text-based localization results on COCO. We compare with the single- and multi-task GPV [20].

Models	Multitask	Image Size	AP50
FRCNN [67,20]	✗	640	75.2
GPV [20]	✓	640	73.0
FindIt-LOC	✗	640	77.9
FindIt-MIX	✓	640	78.6
FindIt	✓	640	**79.7 ± 0.1**

(b) Detection results on COCO. We compare with the single- and multi-task baselines from [27,18]. .

Models	Multitask	Image Size	mAP
FRCNN [18]	✗	1333	37.9
UNiT [27]	✗	1333	**40.6**
UNiT [27]	✓	1333	39.0
FindIt-DET[1]	✗	1024	**40.6**
FindIt-MIX	✓	640	38.4
FindIt	✓	640	39.7 ± 0.1

points. Compared to the single-task baselines, FindIt consistently improves the performance by 3-9 points across the RefCOCO splits, showing the benefits of multitask training.

We note that all results in Table 2 only use COCO box annotation and language corpus pretraining [65]. We do not pretrain on vision and language datasets or use the mask annotations in COCO [54]. Existing approaches [8,16,35] obtain SOTA performance on RefCOCO by pretraining on large vision and language datasets [5,57], visual grounding datasets [35,38,61], or graph relationships [88]. Without using external data, FindIt-MIX is on par with or better than the SOTA method [35] pre-trained with more visual grounding data. Our best-performing model on REC uses a smaller image size (384) than the rest of the paper (640).

To avoid contamination for FindIt, we remove the overlapping images of the RefCOCO val/test sets from the training sets of LOC and DET based on the RefCOCO split they are trained with. For FindIt-MIX, we carefully remove the overlapping images of all RefCOCO val/test sets from all REC, LOC, and DET training sets. The mixing ratio for FindIt-MIX is 2:2:1:1:1:1 among DET:LOC:REC:REC+:REC-g:REC-umd. The FindIt and FindIt-MIX models in Table 2 and Table 3 are the same without task-specific fine-tuning.

Table 3a compares our work on the text-based localization (LOC) task. We compare to the recent GPV method [20] which is is the best approach on this task. For FindIt, we report the mean and standard deviation over four individual RefCOCO splits. FindIt outperforms GPV in all settings. Following GPV [20], we train both LOC and DET tasks on COCO'14 train split (80k images) and report performance on COCO'14 val split (40k images) in Table 3. Table 3b shows our results on detection. We see that our approach is comparable to the full UNiT [27], which uses a detection-specific task head, larger image size, and more training images (COCO'17 vs our COCO'14). Compared to the single-task setup, FindIt shows a similar performance gap to that seen in UNiT's multitask setup. Figure 3 shows examples of FindIt on all three tasks.

[1] FindIt-DET is trained and tested on COCO 17' to match the settings of [27,18].

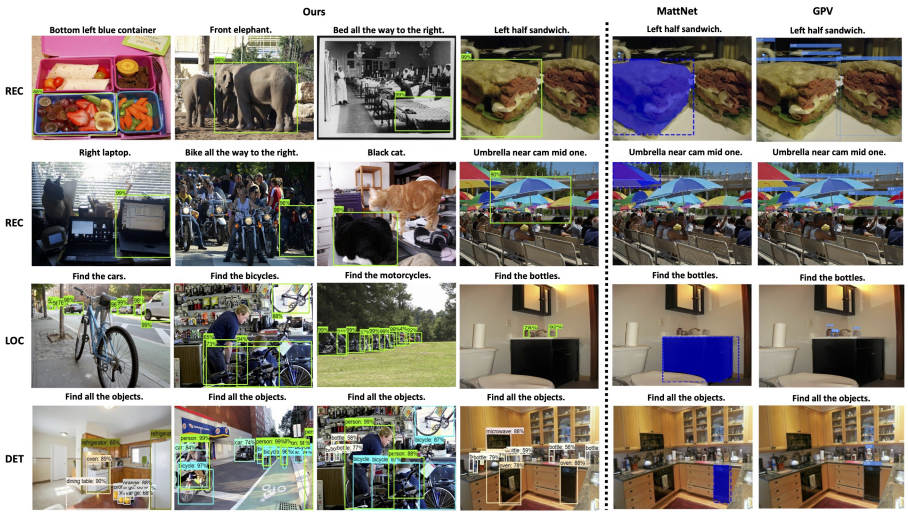

Fig. 3: Visualization of FindIt on REC, LOC, and DET. Compared to existing baselines [89,20], FindIt can perform these tasks well and in a single model.

Table 4: Generalization study through text-based localization task.

(a) Generalization to novel categories.

Model	REC	Base-80	Novel-285	All
FindIt-REC	79.5	21.3	5.2	13.1
FindIt-LOC	-	56.7	15.2	33.9
FindIt-MIX	**84.9**	**57.8**	**18.7**	**36.4**

(b) Generalization to super-categories.

Model	COCO	COCO-O365	O365
FindIt-REC	33.0	18.6	11.0
FindIt-LOC	45.8	25.3	15.3
FindIt-MIX	**49.5**	**30.1**	**17.5**

4.2 Generalization Capabilities of FindIt

We now evaluate the generalization capabilities of the FindIt model presented in Section 4.1. The Objects365 dataset [71] is chosen for the study, because it is independently collected and represents OOD (Out-of-Distribution) data. In addition, the dataset is large, well-annotated with high recall, and contains all of 80 COCO categories and 285 novel categories (365 in total) to assess the generalization of FindIt models. Our models acquire the linguistic knowledge of novel categories from multi-task cross-attention learning and language pretraining [65]. However, as all of our single- and multi-task models share the same language pretraining, the main differences arise from multi-task learning.

Localization on Novel Categories. Even though referring expression models are able to effectively localize objects from complex queries, we want to investigate whether they are able to handle the text-based localization task. Thus, we evaluate the single-task FindIt-REC, FindIt-LOC, and unified training FindIt-MIX models on Objects365 dataset. All FindIt models are identical to their counterparts in Table 2 and Table 3a without further fine-tuning. The

Table 5: Runtime benchmark with recent REC approaches.

Models	Image Size	Backbone	Runtime (ms)
MattNet [89]	1000	R101	378
FAOA [87]	256	DarkNet-53	39
MCN [54]	416	DarkNet-53	56
TransVG [12]	640	R50	62
FindIt	640	R50	107
FindIt	384	R50	57

FindIt-REC model was trained on the RefCOCO UNC split. Table 4a shows the results, where the column "Base-80" evaluates the 80 COCO categories; "Novel-285" evaluates the 285 non-COCO categories; "All" evaluates all 365 categories; "REC" is the performance on RefCOCO UNC. We first observe that FindIt-REC struggles on this task , despite having strong performance on REC. FindIt-LOC model performs much better because it was directly trained for this task. Compared to FindIt-LOC and FindIt-REC, FindIt generalizes better especially on the novel categories of Objects365, because it has acquired broader knowledge about objects and grounding texts through multi-task learning.

Localization on Super-categories. By accepting text inputs, FindIt model relaxes the requirement for a pre-defined set of classes for localization and can generalize beyond the training vocabulary (i.e. COCO categories). We study this behavior by testing on COCO and Objects365 super-categories (e.g. giraffe $\in$ animal, pizza $\in$ food). The setup is identical to Table 4a except that the query category names are replaced with their corresponding super-categories. All models here are the same as in Table 4a. We present the results in Table 4b. The column "COCO" evaluates the COCO super-categories on COCO data; "COCO-O365" evaluates the COCO super-categories on Objects365; "O365" evaluates the Objects365 super-categories on Objects365. Despite the challenging setup, FindIt generalizes better than single-task baselines by a clear margin, showing the merits of broader grounding knowledge provided by multitask learning (see Figure 1 for more examples).

4.3 Analysis and Ablations

Inference Time. We benchmark the inference times across image sizes in Table 5 on the REC task. FindIt is efficient and comparable with existing approaches, while achieving higher accuracy (See Table 2). For fair comparison, all running times are measured on one GTX 1080Ti GPU. Compared to the two-step approach [89], FindIt is more efficient because it trains end-to-end without a need for pre-computed detections.

Task Prompts. We conducted the ablations on the prompts of LOC and DET tasks in Table 6a and found the prompts have minimal effects on performance. Our LOC prompt "Find the X" is one of the prompts used by GPV [20].

Language Model Size. We conducted ablations on the language model sizes in Table 6b and found that larger models are only marginally better. In Table 6b,

REC is the mean performance over all RefCOCO splits. We choose T5 base [65] as the best trade-off between performance and speed.

Multi-level Fusion Architecture. We conduct ablation studies on the fusion architecture and multitask mixing ratios. All experiments of this section are run with a 6x shorter schedule and weaker data augmentation for faster convergence. In all tables we use RefCOCO-UNC as a representative split to evaluate the REC task except for the ablation on language model size. In Table 6c we study the effect of architecture choices on the downstream tasks. We find that attention-based fusion outperforms other alternatives given the same configuration (e.g. 256 dim, 3 layers). In addition, increasing the number of attention layers and the embedding dimension both improve the performance on referring expression, but not as much on detection and localization. We explore multi-scale fusion in Table 6d, and find that using more levels is beneficial for all model sizes we study. Thus, levels (4, 5) are chosen for all experiments. Table 6e delves deeper to show the benefits of attention fusion for REC tasks. With adequate model capacity (e.g. 256 dim, 3 layers), attention fusion outperforms the other alternatives under the same configuration. On the split with the most complex queries (RefCOCO-g), we notice attention fusion performs substantially better than other alternatives. From these studies, we choose (Attention, 256 dim, 3 layers) as our model, because we find the larger alternative (Attention, 512 dim, 6 layers) to perform only marginally better with full training schedule.

Table 6f studies the sampling weight in multitask learning. We find that a simple 1:1:1 ratio achieves a good balance between DET, LOC and REC task performance. Increasing the sampling rate for one task tends to improve the performance at the expense of other tasks. We note that the mixing ratios can be further optimized to improve the performance of any constituent task. We use 256-dimension fusion features, 3 layers, and fusion levels (4, 5) in this ablation.

5 Conclusion

We present Findit, which unifies referring expression comprehension, text-based localization, and object detection tasks. We propose multi-scale cross-attention to unify the disparate localization requirements of these tasks. Without any task-specific design, FindIt surpasses the state of the art on referring expression and text-based localization, shows competitive performance on detection, and generalizes better to out-of-distribution data and novel classes. All of these are accomplished in a single, unified and efficient model.

Acknowledgements. We would like to thank Ashish Vaswani, Prajit Ramachandran, Niki Parmar, David Luan, Tsung-Yi Lin, and other colleagues at Google Research for their advice and helpful discussion.

Table 6: Ablations on task prompts, language model sizes, multi-level fusion architecture design, and mixing ratios.

(a) Ablations on task prompts. The first row corresponds to default FindIt.

LOC Prompt	DET Prompt	LOC	DET
"Find the X"	"Find all the objects"	78.78	38.96
"X"	"This is detection task"	79.13	38.88
GPV [20]	"Find all the objects"	78.92	38.97

(b) Ablations on language model sizes.

Language Model	DET	LOC	REC
T5-Small [65]	38.7	78.7	80.7
T5-Base [65]	38.4	78.7	81.0
T5-Large [65]	38.8	78.8	81.2

(c) Ablations on the fusion mechanism, feature dimension and the number of transformer layers.

Fusion	Dim.	# Layers	DET	LOC	REC
Concat	256	1	35.6	76.6	77.7
Product	256	1	35.4	76.6	78.9
Product	256	3	35.1	76.2	76.7
Attention	128	1	35.7	75.8	75.2
Attention	256	3	35.6	76.5	79.3
Attention	512	6	35.7	76.9	79.3
Attention	1024	12	35.7	77.1	82.1

(d) Ablations on the fusion levels, feature dimension dimensionsion and the number of transformer layers.

Levels	Dim.	# Layers	DET	LOC	REC
(5,)	256	3	35.6	76.7	78.8
(5,)	512	6	34.6	76.0	80.0
(5,)	1024	12	33.0	75.0	80.5
(4, 5)	256	3	35.6	76.5	79.3
(4, 5)	512	6	35.7	76.9	79.3
(4, 5)	1024	12	35.7	77.1	82.1

(e) Ablations on the fusion architecture for the REC tasks.

Fusion	Dim.	Layers	UNC	Plus	G	UMD
Concat.	128	1	76.1	60.4	53.2	62.4
Concat.	256	3	76.8	61.7	54.5	63.6
Concat.	512	6	77.2	61.8	57.1	64.6
Product	128	1	66.5	62.4	55.3	63.3
Product	256	3	76.0	60.9	54.6	62.6
Product	512	6	75.6	60.1	57.4	62.4
Attention	128	1	73.7	57.1	53.9	60.6
Attention	256	3	78.6	62.9	60.8	64.4
Attention	512	6	78.6	65.6	60.6	67.3

(f) Ablations on multitask mixing ratios for all tasks.

DET : LOC : REC	DET	LOC	REC
1 : 1 : 1	35.5	76.6	78.5
2 : 1 : 1	35.9	75.9	77.7
1 : 2 : 1	34.9	77.0	78.3
2 : 2 : 1	35.8	76.8	76.9

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR (2018) 4
2. Bansal, A., Sikka, K., Sharma, G., Chellappa, R., Divakaran, A.: Zero-shot object detection. In: ECCV (2018) 3
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: https://arxiv.org/abs/2005.12872 (2020) 5, 6
4. Changpinyo, S., Pont-Tuset, J., Ferrari, V., Soricut, R.: Telling the what while pointing to the where: Multimodal queries for image retrieval. In: Arxiv: 2102.04980 (2021) 4
5. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: CVPR (2021) 4, 10
6. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollar, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. In: Arxiv: https://arxiv.org/abs/1504.00325 (2015) 4
7. Chen, X., Ma, L., Chen, J., Jie, Z., Liu, W., Luo, J.: Real-time referring expression comprehension by single-stage grounding network. arXiv preprint arXiv:1812.03426 (2018) 3, 9
8. Chen, Y.C., Li, L., Yu, L., Kholy, A.E., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: ECCV (2020) 4, 9, 10
9. Chen, Z., Wang, P., Ma, L., Wong, K.Y.K., Wu, Q.: Cops-ref: A new dataset and task on compositional referring expression comprehension. In: CVPR (2020) 3
10. Cho, J., Lei, J., Tan, H., Bansal, M.: Unifying vision-and-language tasks via text generation. In: 2102.02779 (2021) 4
11. Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J.M.F., Parikh, D., Batra, D.: Visual dialog. In: CVPR (2017) 4
12. Deng, J., Yang, Z., Chen, T., Zhou, W., Li, H.: Transvg: End-to-end visual grounding with transformers. ICCV (2021) 3, 4, 5, 6, 7, 8, 9, 12
13. Desai, K., Johnson, J.: Virtex: Learning visual representations from textual annotations. In: CVPR (2021) 4
14. Dhamija, A.R., Gunther, M., Ventura, J., Boult, T.E.: The overlooked elephant of object detection: Open set. In: WACV (2020) 3
15. Gan, C., Li, Y., Li, H., Sun, C., Gong, B.: Vqs: Linking segmentations to questions and answers for supervised attention in vqa and question-focused semantic segmentation. In: ICCV (2017) 4
16. Gan, Z., Chen, Y.C., Li, L., Zhu, C., Cheng, Y., Liu, J.: Large-scale adversarial training for vision-and-language representation learning. In: NeurIPS (2020) 4, 9, 10
17. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR (2014) 1, 3
18. Girshick, R., Radosavovic, I., Gkioxari, G., Dollár, P., He, K.: Detectron. https://github.com/facebookresearch/detectron (2018) 5, 8, 10
19. Gu, X., Lin, T., Kuo, W., Cui, Y.: Zero-shot detection via vision and language knowledge distillation. CoRR **abs/2104.13921** (2021), https://arxiv.org/abs/2104.13921 3

20. Gupta, T., Kamath, A., Kembhavi, A., Hoiem2, D.: Towards general purpose vision systems. In: arxiv.org/abs/2104.00743 (2021) 2, 3, 4, 5, 7, 8, 10, 11, 12, 14
21. Gupta, T., Vahdat, A., Chechik, G., Yang, X., Kautz, J., Hoiem, D.: Contrastive learning for weakly supervised phrase grounding. In: ECCV (2020) 1
22. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask r-cnn. In: ICCV (2017) 3, 5, 7
23. Hinami, R., Satoh, S.: Discriminative learning of open-vocabulary object retrieval and localization by negative phrase augmentation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (2018) 2, 4
24. Hong, R., Liu, D., Mo, X., He, X., Zhang, H.: Learning to compose and reason with language tree structures for visual grounding. TPAMI (2019) 9
25. Hu, Ronghang, R.M.D.T.: Segmentation from natural language expressions. european conference on computer vision. In: ECCV (2016) 1
26. Hu, R., Rohrbach, M., Andreas, J., Darrell, T., Saenko, K.: Modeling relationships in referential expressions with compositional modular networks. In: CVPR. pp. 1115–1124 (2017) 3, 9
27. Hu, R., Singh, A.: Unit: Multimodal multitask learning with a unified transformer. In: arxiv.org/abs/2102.10772 (2021) 4, 10
28. Huang, G., , B.P., Zhu, Z., Rivera, C., Soricut, R.: Multimodal pretraining for dense video captioning. In: AACL-IJCNLP (2020) 4
29. Huang, Z., Zeng, Z., Huang, Y., Liu, B., Fu, D., Fu, J.: Seeing out of the box: End-to-end pre-training for vision-language representation learning. In: CVPR (2021) 4
30. Hudson, D.A., Manning, C.D.: Gqa: a new dataset for compositional question answering over realworld images. In: CVPR (2019) 3
31. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q.V., Sung, Y., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML (2021) 4
32. Jiang, H., Misra, I., Rohrbach, M., Learned-Miller, E., Chen, X.: In defense of grid features for visual question answering. In: CVPR (2020) 4
33. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C.L., Girshick, R.: A diagnostic dataset for compositional language and elementary visual reasoning. In: CVPR (2017) 3
34. Jun Xu, Tao Mei, T.Y., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: CVPR (2016) 4
35. Kamath, A., Singh, M., LeCun, Y., Misra, I., Synnaeve, G., Carion, N.: Mdetr - modulated detection for end-to-end multi-modal understanding. In: https://arxiv.org/abs/2104.12763 (2021) 3, 4, 9, 10
36. Kant, Y., Moudgil, A., Batra, D., Parikh, D., Agrawal, H.: Contrast and classify: Training robust vqa models. In: ICCV (2021) 4
37. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: ICML (2021) 4
38. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., Bernstein, M., Fei-Fei, L.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. In: https://arxiv.org/abs/1602.07332 (2016) 4, 10
39. Kudo, T., Richardson, J.: Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. arXiv preprint arXiv:1808.06226 (2018) 8
40. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. In: Arxiv:https://arxiv.org/abs/1908.03557 (2019) 4

41. Li*, L.H., Zhang*, P., Zhang*, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., Chang, K.W., Gao, J.: Grounded language-image pre-training. In: CVPR (2022) 4
42. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y., Gao, J.: Oscar: Object-semantics aligned pre-training for vision-language tasks. In: ECCV (2020) 4
43. Liao, Y., Liu, S., Li, G., Wang, F., Chen, Y., Qian, C., Li, B.: A real-time cross-modality correlation filtering method for referring expression comprehension. In: CVPR. pp. 10880–10889 (2020) 3, 9
44. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017) 6, 8
45. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017) 3, 5, 6
46. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV (2014) 3, 4, 5, 8
47. Lin, X., Bertasius, G., Wang, J., Chang, S.F., Parikh, D.: Vx2text: End-to-end learning of video-based text generation from multimodal inputs. In: CVPR (2021) 4
48. Liu, C., Lin, Z., Shen, X., Yang, J., Lu, X., Yuille, A.: Recurrent multimodal interaction for referring image segmentation. In: ICCV (2017) 4
49. Liu, D., Zhang, H., Wu, F., Zha, Z.J.: Learning to assemble neural module tree networks for visual grounding. In: ICCV. pp. 4673–4682 (2019) 3, 9
50. Liu, X., Wang, Z., Shao, J., Wang, X., Li, H.: Improving referring expression grounding with cross-modal attention-guided erasing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1950–1959 (2019) 3, 8, 9
51. Liu, Z., Stent, S., Li, J., Gideon, J., Han, S.: Loctex: Learning data-efficient visual representations from localized textual supervision. In: ICCV (2021) 4
52. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: CVPR (2019) 4
53. Lu, J., Goswami, V., Rohrbach, M., Parikh, D., Lee, S.: 12-in-1: Multi-task vision and language representation learning. In: CVPR (2020) 4
54. Luo, G., Zhou, Y., Sun, X., Cao, L., Wu, C., Deng, C., Ji, R.: Multi-task collaborative network for joint referring expression comprehension and segmentation. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 10031–10040 (2020) 3, 4, 8, 9, 10, 12
55. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: CVPR (2016) 1, 3, 9
56. Margffoy-Tuay, E., Perez, J.C., Botero, E., Arbelaez, P.: Dynamic multimodal instance segmentation guided by natural language queries. In: ECCV (2018) 1
57. Ordonez, V., Kulkarni, G., Berg, T.L.: Im2text: Describing images using 1 million captioned photographs. In: NeurIPS (2011) 10
58. Papageorgiou, C., Oren, M., Poggio, T.: A general framework for object detection. In: ICCV (1998) 1
59. Peng, J., Bu, X., Sun, M., Zhang, Z., Tan, T., Yan, J.: Large-scale object detection in the wild from imbalanced multi-labels. In: CVPR (2020) 5

60. Plummer, B.A., Shih, K.J., Li, Y., Xu, K., Lazebnik, S., Sclaroff, S., Saenko, K.: Revisiting image-language networks for open-ended phrase detection. In: TPAMI (2020) 1, 4
61. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE international conference on computer vision. pp. 2641–2649 (2015) 10
62. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: International Journal of Computer Vision (2017) 1, 3
63. Qiao, Y., Deng, C., Wu, Q.: Referring expression comprehension: A survey of methods and datasets. In: IEEE TMM (2020) 3
64. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021) 4
65. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. In: Journal of Machine Learning Research (JMLR) (2020) 6, 8, 10, 11, 13, 14
66. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: CVPR (2016) 1, 3, 5, 6
67. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (2015) 2, 3, 5, 6, 8, 10
68. Ronghang Hu, Huazhe Xu, M.R.J.F.K.S.T.D.: Natural language object retrieval. In: CVPR (2016) 1
69. Rowley, H., Baluja, S., Kanade, T.: Human face detection in visual scenes. In: Advances in Neural Information Processing Systems (1995) 1
70. Sahar Kazemzadeh, Vicente Ordonez, M.M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: EMNLP (2014) 1, 3
71. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Li, J., Zhang, X., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: ICCV (2019) 3, 11
72. Srinivasan, K., Raman, K., Chen, J., Bendersky, M., Najork, M.: Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In: arXiv:2103.01913 (2021) 4
73. Suhr, A., Lewis, M., Yeh, J., Artzi, Y.: A corpus of natural language for visual reasoning. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (2017) 4
74. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. In: EMNLP (2019) 4
75. Vaillant, R., Monrocq, C., Cun, Y.L.: An original approach for the localization of objects in images. In: IEEE Proc. Visual Image Signal Processing (1994) 1
76. Varun K Nagaraja, V.I.M., Davis, L.S.: Modeling context between objects for referring expression understanding. In: ECCV (2016) 3
77. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017) 4, 6
78. Viola, P., Jones, M.: Robust real-time object detection. In: International Journal of Computer Vision (2001) 1

79. Wang, L., Li, Y., Huang, J., Lazebnik, S.: Learning two-branch neural networks for image-text matching tasks. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (2018) 4
80. Wang, M., Azab, M., Kojima, N., Mihalcea, R., Deng, J.: Structured matching for phrase localization. In: ECCV (2016) 1
81. Wang, P., Wu, Q., Cao, J., Shen, C., Gao, L., Hengel, A.v.d.: Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In: CVPR. pp. 1960–1968 (2019) 3, 9
82. Xian, Y., Lampert, C.H., Schiele, B., Akata, Z.: Zero-shot learning – a comprehensive evaluation of the good, the bad and the ugly. In: TPAMI (2018) 3
83. Xie, N., Lai, F., Doran, D., Kadav, A.: Visual entailment: A novel task for fine-grained image understanding. In: https://arxiv.org/abs/1901.06706 (2019) 4
84. Xu, H., Ming Yan, Chenliang Li, B.B.S.H.W.X., Huang, F.: E2e-vlp: End-to-end vision-language pre-training enhanced by visual learning. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (2021) 4
85. Yang, S., Li, G., Yu, Y.: Dynamic graph attention for referring expression comprehension. In: ICCV. pp. 4644–4653 (2019) 3, 9
86. Yang, Z., Chen, T., Wang, L., Luo, J.: Improving one-stage visual grounding by recursive sub-query construction. In: ECCV (2020) 3, 4, 7, 8, 9
87. Yang, Z., Gong, B., Wang, L., Huang, W., Yu, D., Luo, J.: A fast and accurate one-stage approach to visual grounding. In: ICCV. pp. 4683–4693 (2019) 2, 3, 4, 5, 6, 7, 8, 9, 12
88. Yu, F., Tang, J., Yin, W., Sun, Y., Tian, H., Wu, H., Wang, H.: Ernie-vil: Knowledge enhanced vision-language representations through scene graph. In: AAAI (2021) 4, 10
89. Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., Berg, T.L.: Mattnet: Modular attention network for referring expression comprehension. In: CVPR (2018) 2, 3, 4, 8, 9, 11, 12
90. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: ECCV (2016) 1, 3, 8, 9
91. Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. In: CVPR (June 2019) 4
92. Zhang, H., Niu, Y., Chang, S.F.: Grounding referring expressions in images by variational context. In: CVPR. pp. 4158–4166 (2018) 3, 8, 9
93. Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., Gao, J.: Vinvl: Revisiting visual representations in vision-language models. In: CVPR (2021) 4
94. Zhao, F., Li, J., Zhao, J., Feng, J.: Weakly supervised phrase localization with multi-scale anchored transformer network. In: CVPR (2018) 4
95. Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J.J., Gao, J.: Unified vision-language pre-training for image captioning and vqa. In: AAAI (2020) 4
96. Zhu, P., Wang, H., Saligrama, V.: Zero-shot detection. In: IEEE Transactions on Circuits and Systems for Video Technology (2018) 3
97. Zhuang, B., Wu, Q., Shen, C., Reid, I., van den Hengel, A.: Parallel attention: A unified framework for visual object discovery through dialogs and queries. In: CVPR. pp. 4252–4261 (2018) 9