Who's Waldo? Linking People Across Text and Images

Claire Yuqing Cui[1*] Apoorv Khandelwal[1*] Yoav Artzi[1,2] Noah Snavely[1,2] Hadar Averbuch-Elor[1,2]
[1]Cornell University [2]Cornell Tech
{yc2296, ak2254, yoavartzi, snavely, hadarelor}@cornell.edu

Abstract

We present a task and benchmark dataset for ***person-centric visual grounding****, the problem of linking between people named in a caption and people pictured in an image. In contrast to prior work in visual grounding, which is predominantly object-based, our new task masks out the names of people in captions in order to encourage methods trained on such image–caption pairs to focus on contextual cues, such as the rich interactions between multiple people, rather than learning associations between names and appearances. To facilitate this task, we introduce a new dataset,* **Who's Waldo**, *mined automatically from image–caption data on Wikimedia Commons. We propose a Transformer-based method that outperforms several strong baselines on this task, and release our data to the research community to spur work on contextual models that consider both vision and language. Code and data are available at:* `https://whoswaldo.github.io`

Figure 1. By studying this picture and caption, we can use contextual cues to link between the people referred to in the text and their visual counterparts, even if we are unfamiliar with the specific individuals. This capability requires understanding of a broad set of interactions (*e.g.* "passing") and expected behaviors (*e.g.* players pass to their teammates). We propose the task of *person-centric visual grounding*, where we abstract over identity names (*e.g.* masking out *Sam Schulz* and *Curtly Hampton* with `[NAME]` tokens) to encourage algorithms to emulate such contextual reasoning.

1. Introduction

The correspondence between people observed in images and their mentions in text is informed by more than simply their identities and our knowledge of their appearances. Consider the image and caption in Figure 1. We often see such image–caption pairs in newspapers and, as humans, are skilled at recovering associations between the people depicted in images and their references in captions, even if we're unfamiliar with the specific people mentioned. This ability requires complex visual reasoning skills. For the example in Figure 1, we must understand an underlying activity ("passing") and determine who is passing the ball, who is being passed to, and which people in the image are not mentioned at all.

In this paper, we present a person-centric vision-and-language grounding task and benchmark. The general problem of linking between textual descriptions and image regions is known as *visual grounding*, and is a fundamental capability in visual semantic tasks with applications including image captioning [66, 41, 3], visual question answering [19, 20, 26] and instruction following [4, 43, 7]. Our task and data depart from most existing works along two axes. First, our task abstracts over identity information, instead focusing specifically on the relations and properties specified in images and text. Second, rather than using data annotated by crowd workers, we leverage captions originating from real-life data sources.

While visual grounding has traditionally centered around localizing objects based on referring expressions, we observe that inferring associations based on expressions in person-centric samples—*i.e.* people's names—could lead to problematic biases (*e.g.* with regards to gender). Hence, we formulate the task to use captions that mask out people's names. This allows for an emphasized focus on the context—both in image and text—where the person appears, requiring models to understand complex asymmetric human interactions and expected behaviors. For instance, in the example in Figure 1 we might expect a player to pass to someone on their own team.

* Equal contribution

To explore this problem, we create *Who's Waldo*: a collection of nearly 300K images of people paired with textual descriptions and automatically annotated with alignments between mentions of people's names and their corresponding visual regions. *Who's Waldo* is constructed from the massive public catalog of freely-licensed images and descriptions in Wikimedia Commons. We leverage this unique data source to automatically extract image–text correspondences for over 200K people. We also provide evaluation sets that are validated using Amazon Mechanical Turk and demonstrate that our annotation scheme is highly accurate.

To link people across text and images, we propose a Transformer-based model, building upon recent work on learning joint contextualized image–text representations. We use similarity measures in the joint embedding space between mentions of people and image regions depicting people to estimate these links. The contextualized Transformer-based representations are particularly suited to handle the masked names, by shifting the reasoning to surrounding contextual cues such as verbs indicating actions and adjectives describing visual qualities. Our results demonstrate that our model effectively distinguishes between different individuals in a wide variety of scenes that capture complex interactions, significantly improving over strong baselines.

2. Related Work

Visual Grounding. The goal of visual grounding is to localize objects in an image given a textual description. Tasks are typically formulated to either recover correspondences between object region proposals and text, or compute attention maps over the whole image. Referring expression comprehension (REC) is a common variant of this problem, where the goal is to identify an image region corresponding to a sentential description (*e.g.* [51, 17, 14, 67, 37]). Sadhu et al. [52] recently extended this task to a zero-shot setting that also considers expressions with unseen nouns. Qiao et al. [49] provide a comprehensive survey on REC.

Still, this line of work has made limited use of descriptions of relationships between objects. The Flickr30K Entities dataset [47] opened up new avenues for modeling such dependencies by including images, full captions, and ground-truth links between regions and phrases for nearly a hundred object categories. Several methods have since been proposed to visually ground objects from textual descriptions that describe multiple objects [61, 46, 38]. A weakly-supervised setting, which assumes that ground truth links between regions and phrases are not available, has also gained attention, with discriminative and contrastive objectives [64, 62, 24], visual and linguistic consistencies [8] and multilevel aggregation strategies [70, 1, 13] used to align the image and language spaces.

However, most existing tasks in visual grounding permit models to reason over referring expressions directly (allowing models to learn priors over different object categories). Our proposed task instead requires models to exclusively reason over context and interactions between objects, as the referring expressions (*i.e.* names) are masked.

The creation of most datasets related to visual grounding involves a time-consuming, expensive annotation process that includes both (i) generating referring expressions or full textual descriptions for a given image, and (ii) annotating corresponding regions in the image (*e.g.* [36, 28, 47, 68]). We instead construct *Who's Waldo* through an automatic approach inspired by Conceptual Captions [56]. While that work uses alt-text image descriptions from HTML (which are noisy and must be aggressively filtered), we use raw descriptions obtained from captions in Wikimedia Commons.

People-centric Tasks. Person identification [6, 32, 71] is a task related to the one we propose, and is formulated as a comparison between reference and target images, aiming to determine whether these belong to the same identity. Our work instead focuses on learning a contextual correspondence between image regions and textual captions describing people and their depicted interactions. For ethical reasons (see Ethical Considerations in Section 4), our released dataset does not contain identity information, and thus cannot be easily modified to train such models.

Another related people-centric task is to select a set of attributes that will generate a description for each person in an image that distinguishes that individual from others in that image [53]. Finally, Aneja *et al.* detect out-of-context image and caption pairs, using a dataset collected from news and fact-checking websites. Their data (specifically, the subset of images capturing people) could be used to augment ours.

Task-agnostic Joint Image-Text Representations. Recent advances have led to a surge of interest in task-agnostic joint visual and textual representations [39, 59, 34, 57, 10, 58, 72, 40, 35, 21]. Several works, such as LXMERT [59] and ViLBERT [39] learn these representations using two-stream transformers [60] (one per modality). Others, including VisualBERT [34], VL-BERT [57] and UNITER [10], use a unified architecture. In our work, we leverage these task-agnostic features to learn to link between the individuals described in the text and their visual counterparts.

3. Person-centric Visual Grounding

Given an image I with $m \geq 1$ people detections and a corresponding caption x_s referring to $n \geq 1$ people (with each person mentioned one or more times), we wish to find a mapping from referred people to visual detections.

We expect to produce a partial, injective (one-to-one) mapping, since not all referred people will be pictured and no two referred people should map to the same detection. We also find that this mapping is not necessarily surjective (onto), since the image may picture people who are not named in the caption and there could exist detections not mapped to by any referred people.

In-the-wild captions featuring people often refer to them by name. However, reasoning about visual grounding using actual names of people involves two challenges: the diversity of names creates significant data sparsity, and their surface form (*i.e.*, the text itself) elicits strong biases, e.g., with regard to gender. We therefore abstract over the surface form of the names by replacing each name with the placeholder token `[NAME]`. This encourages models to focus on the textual context of the names, including adjectives and adverbs that hint at the person's visual appearance and verbs that indicate the action they partake in. In other words, by masking names we seek models that do not memorize what specific people look like, or form stereotypical associations based on specific names, but must instead learn richer contextual cues. As part of our dataset, we provide a mapping from referred people to their respective sets of referring `[NAME]` tokens.

While visual grounding has traditionally centered around localization of objects (including unnamed people), we find that visual grounding in the context of named people (which we denote as *person-centric*) presents additional opportunities. In object-centric visual grounding, referring expressions are not masked, allowing models to also learn by matching images and object classes, rather than entirely from context. Moreover, data for our task (*i.e.*, captioned images of people) is readily available on the web and matches a realistic distribution more closely than object datasets, whose pairs are annotated by workers for the sole purpose of visual grounding tasks.

Evaluation. Given a mapping produced by an algorithm for an input example, we evaluate by computing accuracy against ground truth links of referred people and detections. This is unlike prior works that extract hundreds of candidate boxes and approximate correct matches using either intersection-over-union ratios or the pointing game, which requires the model to predict a single point per phrase. We also enforce that the people in test images and captions do not appear during training.

4. The *Who's Waldo* Dataset 

In this section, we describe *Who's Waldo*[1], a new dataset with 270K image–caption pairs, derived from Wikimedia Commons.[2] We first describe the process of constructing and annotating this dataset, then present an analysis over dataset statistics. We show samples from our dataset, along with their annotations, in Figure 2.

Data Collection. Under the broader "People by name" category in Wikimedia Commons there are 407K categories named after people, each with their own hierarchy of subcategories. We refer to this set of people as *Wikimedia identities*. We identified all sub-categories that are person-centric (*e.g*. "Barack Obama playing basketball" or "Sally Ride on Challenger in 1983", rather than "John F. Kennedy International Airport") by tokenizing names, matching tokens with regular expressions, and tagging parts of speech. We then downloaded 3.5M images, collated duplicates, and retained references to the Wikimedia identities they originate from. We observe that images originating from an identity are very likely to depict that identity.

Many images on Wikimedia Commons are also associated with human-provided English captions that describe these images by naming the people present and detailing their settings and interactions. We collected these captions and pre-processed them by pattern matching with regular expressions to remove Wikimedia-specific text structures. We also removed phrases that are variants of "photo by [photographer name]", since photographers are often named in captions but are not pictured in images.

Detecting People in Images and Captions. To detect bounding boxes for people in images, we used a Switchable Atrous Convolution model with a Cascade R-CNN and ResNet-50 backbone from MMDetection [48, 9] trained on COCO [36]. We subsequently estimated 133 whole-body keypoints using a top-down DarkPose model from MMPose [69, 11] (trained on COCO [36] and finetuned on COCO-WholeBody [27]).

We applied a pre-trained Punkt sentence tokenizer from NLTK [29, 5] to all captions and performed named entity recognition on each sentence using FLAIR [2] to identify person names. We observe that people can be mentioned more than once in captions and without exact matches (*e.g*., as "William" and "Bill", or "Barack" and "Obama"). Therefore, we used neural coreference resolution models from AllenNLP [33, 22] to cluster multiple name entities as individual referred persons.

Estimating Ground Truth Links. To produce supervision for our task, we automatically generated ground truth links from referred people in captions to detections of people in images. As we will describe, Wikimedia Commons provides reference faces for many referred people. As we can also generate face images for our image detections (via face alignment from estimated pose landmarks), we computed a similarity matrix using FaceNet embeddings [55, 54] between reference faces and detected faces. By finding a minimum weight bipartite matching [31] in this matrix and applying a threshold (set empirically to 0.46), we recovered a partial mapping from referred people to detections.

We find reference faces for referred people as follows. First, we associate referred people with Wikimedia identities (via the prior coreference resolution step). We also

[1] Icon created by Stefan Spieler from the Noun Project
[2] `https://commons.wikimedia.org`

Left: "**Justyna Kowalczyk**, **Kikkan Randall** and **Ingvild Flugstad Østberg** at the Royal Palace Sprint, part of the FIS World Cup 2012/2013, in Stockholm on March 20, 2013. **Kikkan Randall** won the sprint cup." **Center**: "**Cheick Diallo** blocks **Allonzo Trier** (#20) in front of **Luke Kennard** (#5) and **Carlon Bragg** (#31) in the 2015 McDonald's All-American Boys Game." **Right**: "At the Gagarin Cosmonaut Training Center in Star City, Russia, Expedition 41/42 backup crew members **Scott Kelly** of NASA (left), **Gennady Padalka** of the Russian Federal Space Agency (Roscosmos, center) and **Mikhail Kornienko** of Roscosmos (right) clasp hands as they pose for pictures in front of a Soyuz simulator at the start of final qualifications."

Figure 2. **Samples from *Who's Waldo***, showing detected named entities in bold and entities linked with image regions in unique colors, corresponding to the boxes on the images. Unmatched boxes and entities are colored in black.

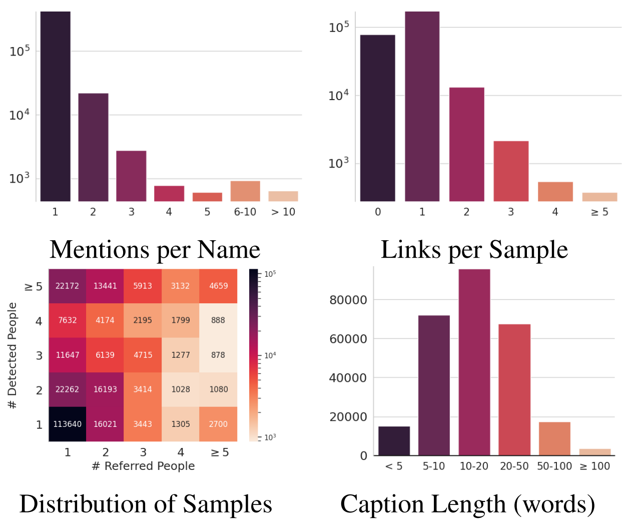


Figure 3. ***Who's Waldo* statistics,** including the number of number of mentions (occurrences) for referred people in captions, ground truth box-name links per sample, distribution of samples and caption length (by words).

find that many Wikimedia identities have *primary* images on Wikimedia Commons, which prominently display their faces. We treat these as reference faces of referred people. However, not all referred people have such associations, so our ground truth links are a subset of all links.

Dataset Size and Splits. The above process yields 271,747 image–caption pairs. Figure 3 summarizes the distributions of annotations and identities present in *Who's Waldo*.

We split these into 179K training, 6.7K validation, and 6.7K test image–caption pairs. We generate the validation and test splits without overlapping identities from training and by ensuring that examples are challenging and correctly annotated. To do so, we first randomly select 16K identities and produce a validation and test superset from examples containing these identities (observing that additional identities likely appear in these examples as well). We generate the training set from all remaining examples that do not contain *any* identities in the superset. We then remove all (trivial) examples from the superset with exactly one person detection and one referred person. We manually validate this superset further as described below and divide the resulting examples into validation and test splits.

Validating Test Images with AMT. While our method approximates ground truth mappings, we want subsets for evaluation that only include correct ground truth links. To that end, we used Amazon Mechanical Turk (AMT) to remove test set examples with incorrect annotations. Given a ground truth link (*i.e.* identity name and image crop of detected person), we defined the following yes/no AMT task: "Does this [detection crop] contain [identity name]?". For ease of comparison, we also provided workers with a reference image and a link to additional photos for that identity. We assign each ground truth link to two workers. Finally, we select all pairs for which both workers answered "yes".

We manually inspected 400 responses and—accounting for worker disagreement and error—estimate that our automatic technique was accurate for approximately 95.5% of links in superset examples. However, after removing any examples for which either worker answered "no", we estimate that over 98.5% of links in the retained examples are accurate. Please refer to the supplemental material for additional visualizations over our dataset and generated links.

Ethical Considerations People-centric datasets pose ethi-

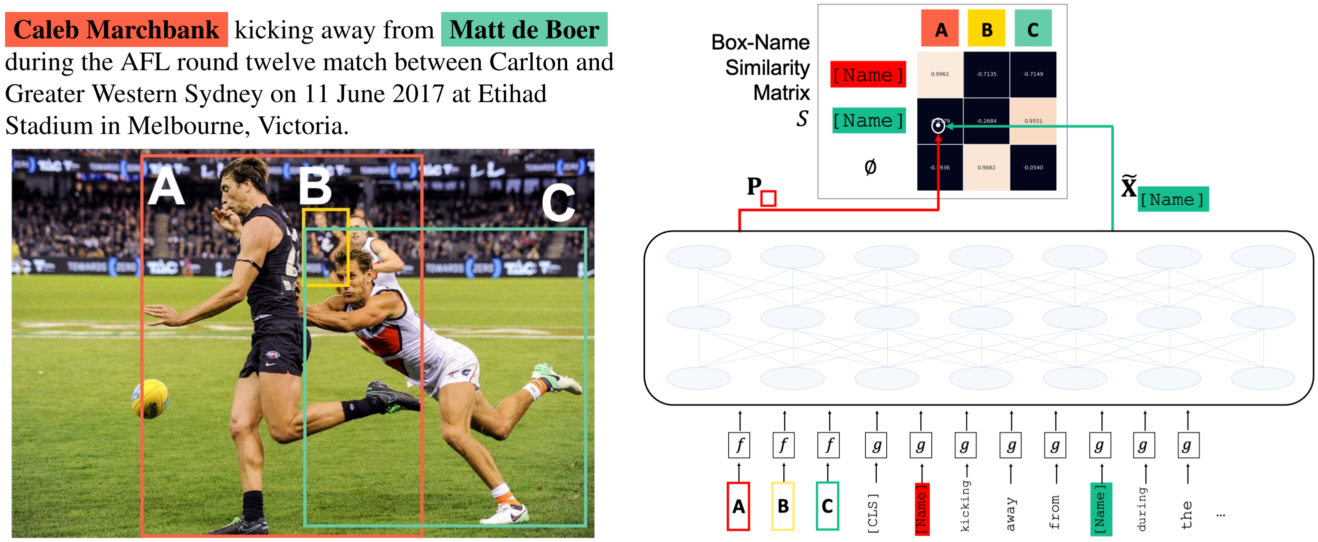


Figure 4. **Overview of our approach.** Features are extracted from image regions and words and combined with a Transformer to learn similarities between people detected in the image (boxes A-C, colored in unique colors) and names mentioned in the caption (Caleb Marchbank and Matt de Boer in the example above). Correspondences are depicted by matching colors.

cal challenges. For example, ImageNet [15] has been scrutinized based on issues inherited from the "person" category in WordNet [12, 65]. Our task and dataset were created with careful attention to ethical questions, which we encountered throughout our work. Access to our dataset will be provided for research purposes only and with restrictions on redistribution. Additionally, as we mask all names in captions, our dataset cannot be easily repurposed for unintended tasks, such as identification of people by name. Due to biases in our data source, we do not consider the data appropriate for developing non-research systems without further processing or augmentation. More details on distribution and intended uses are provided in a supplemental datasheet [23].

5. Method

In this section we present an approach for linking people in text and images. We use a multi-layer Transformer [60] to learn joint image–text representations such that referred people and their corresponding image regions will be highly similar, while those that do not correspond will be dissimilar. For brevity, we refer to the n names of referred people as *names* and m image regions of detected people as *boxes*.

5.1. Model

Our method is based on the recent UNITER Transformer model [10]. As shown in their work, their pretrained model can be leveraged for a wide variety of downstream vision-and-language tasks. In this section, we show how UNITER can be modified for our task and fine-tuned on our dataset. An overview of our approach is shown in Figure 4.

We extract visual features for each person detection p using a fully-convolutional variant of Faster R-CNN [3]. Visual features are concatenated with encodings of their spatial coordinates,[3] yielding spatial-visual features $f(p)$. We tokenize words into WordPieces [63]. In accordance with our task, names are symbolized by `[NAME]` tokens. For each sub-word w, we extract features $g(w)$ that are composed of a token embedding and position embedding.

We feed these spatial-visual and textual features into a Transformer model that uses self-attention layers to learn a contextual representation and captures a more context-specific representation in upper-hidden layers [18]. We denote the final hidden layer of spatial-visual features as $\mathbf{P}_k$ and of textual features as $\mathbf{X}_l$, where $\mathbf{P}_k, \mathbf{X}_l \in \mathbb{R}^{768}$.

From these contextualized representations, we construct a box–name similarity matrix S (top-right of Figure 4). This matrix measures cosine similarities $S_{i,j}$ between the i-th name and the j-th box:

$$S_{i,j} = \frac{\mathbf{P}_j^T \tilde{\mathbf{X}}_i}{\|\mathbf{P}_j\|_2 \|\tilde{\mathbf{X}}_i\|_2}, \tag{1}$$

where $\tilde{\mathbf{X}}_i$ is an embedding averaged over all `[NAME]` tokens for mentions of the i-th referred person in a caption.

During inference, for each referred person, we select its corresponding detection as the most similar box in S.

5.2. Learning

To train our model, we propose the following loss terms that operate on the similarity matrix S: (1) box–name matching losses defined within and across images and (2) an unlinked box classification loss.

[3] Following [10], these are: $[x_1, y_1, x_2, y_2, w, h, w \times h]$.

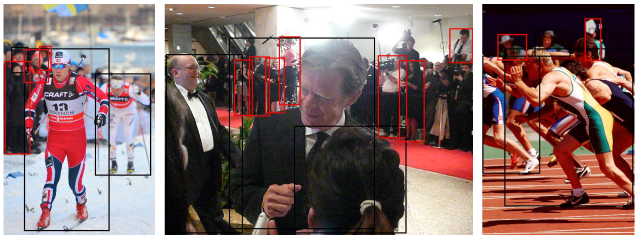


Figure 5. **Selecting unlinked boxes.** We select small and blurry boxes (colored in red) for our proposed classification loss, encouraging the model to focus on larger (and less blurry) people.

Method	Training Data	Accuracy
Full Names		
Gupta et al. [24]	COCO	36.9 ± 1.04
Gupta et al. [24]	Flickr30K Entities	39.3 ± 1.05
SL-CCRF [38]	Flickr30K Entities	43.5 ± 1.06
MAttNet [67]	RefCOCOg	43.6 ± 1.06
UNITER [10]	Multiple [36, 30, 44, 56]	36.3 ± 1.03
Random		
Gupta et al. [24]	COCO	39.3 ± 1.05
Gupta et al. [24]	Flickr30K Entities	41.1 ± 1.06
SL-CCRF [38]	Flickr30K Entities	44.1 ± 1.07
MAttNet [67]	RefCOCOg	44.0 ± 1.07
UNITER [10]	Multiple [36, 30, 44, 56]	38.4 ± 1.04
Constant		
Gupta et al. [24]	COCO	35.6 ± 1.03
Gupta et al. [24]	Flickr30K Entities	38.2 ± 1.04
SL-CCRF [38]	Flickr30K Entities	46.4 ± 1.07
MAttNet [67]	RefCOCOg	24.1 ± 0.92
UNITER [10]	Multiple [36, 30, 44, 56]	34.2 ± 1.02
Random	–	30.9 ± 0.99
Big→Small	–	48.2 ± 1.07
L→R (All)	–	38.4 ± 1.04
L→R (Largest)	–	57.7 ± 1.06
Ours	*Who's Waldo*	**63.5 ± 1.03**

Table 1. Evaluation on the *Who's Waldo* test set. We compare against prior grounding methods using multiple configurations, varying according to how names are processed. We also compare to several simple baselines, detailed in the text.

Box–Name Matching Losses. We define box–name matching losses within images (supervising estimated correspondences with ground truth links) and across images (using a discriminative objective over image–caption pairs).

We compute the estimated probability for a ground truth link (i, j) over different boxes ($p = \text{Softmax}(S_{i,:})_j$) and also over different names ($q = \text{Softmax}(S_{:,j})_i$) in its corresponding image–caption pair. We minimize cross-entropy losses over these for all ground truth links L in a batch:

$$\mathcal{L}_{\mathsf{intra}} = -\frac{1}{|L|} \sum_{l \in L} \left[\log p^{(l)} + \log q^{(l)} \right] \tag{2}$$

Because we would like to leverage additional images during training (*i.e.*, those without ground truth links), we also compute a matching loss across images containing a single box and name (which are likely to represent the same person). We sample positive and negative box–name pairs. Negative pairs are generated by replacing the box with one from another image (and of a different person). We minimize a binary cross-entropy loss $\mathcal{L}_{\mathsf{inter}}$ over these pairs.

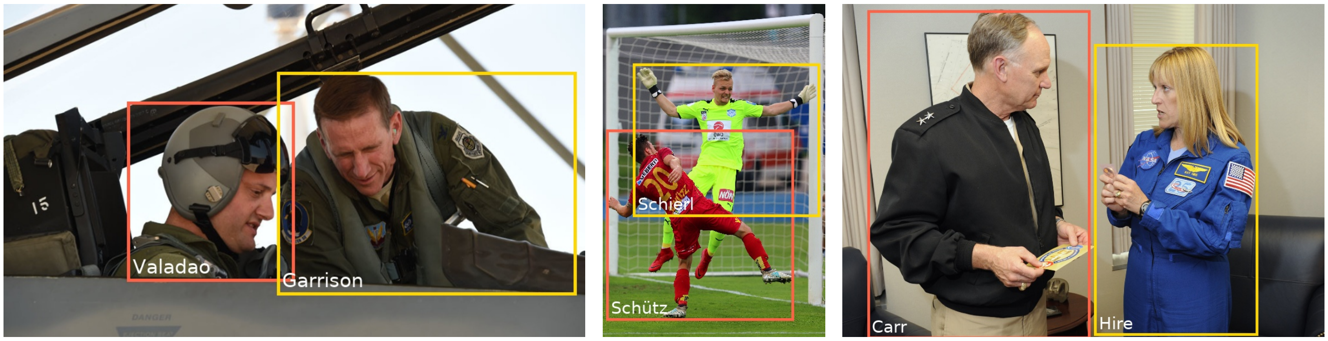
Unlinked Box Classification Loss. As not all people depicted in an image are referred to in its caption, we augment S with a constant null name $\tilde{\mathbf{X}}_{\varnothing}$. We formulate a binary cross-entropy classification loss over similarities between boxes and $\tilde{\mathbf{X}}_{\varnothing}$. We process these similarities $S_{i=\varnothing,j}$ through a sigmoid function to obtain normalized values. Boxes linked to names are considered negative matches (*i.e.*, these should yield low similarities with $\tilde{\mathbf{X}}_{\varnothing}$).

We cannot assume all other boxes are positive matches (*i.e.*, should yield high similarities with $\tilde{\mathbf{X}}_{\varnothing}$) as we are only provided with partial ground truth correspondences from the algorithm in Section 4. Instead, we select unlinked boxes that are (1) *insignificant* compared to other boxes in the image and (2) *blurry*. Both are measured using the a detected person's face (computed from whole body landmarks): a face image f is considered insignificant if $\text{Area}(f) < 0.6 \cdot \text{Area}(f_{\text{largest}})$ and blurry if $\text{Var}(\Delta(f)) < 50$ [45], where f_{largest} is the largest face in the image and Δ is the Laplace operator. Figure 5 shows several images from our dataset with unlinked boxes in red. We minimize a binary cross-entropy loss $\mathcal{L}_{\varnothing}$ over images containing such positive and negative matches.

This loss, in addition to providing us a means of directly estimating whether or not a given box is referred to in a caption, also implicitly encourages the contextualized representations of insignificant and blurry faces to be distinguishable from others. As we show in our results, this improves the accuracy of identifying referred people, allowing the model to focus on more relevant boxes.

6. Results and Evaluation

We compare our model to other visual grounding methods trained on a variety of datasets. We study four key questions: How well do previous methods for visual grounding perform on our proposed task? To what extent is our model reasoning over complex multimodal signals? What is the impact of our design choices? And, what has our model learned? We also present qualitative results (Figure 6 and supplemental material), which highlight the complexity and unique challenges of our proposed task.



U.S. Air Force Colonel **Clay Garrison** goes over some final instructions with U.S. Congressman **David Valadao** prior to take-off from the Fresno Air National Guard Base May 27, 2015.

Daniel Schütz (#20) behind goalkeeper **Domenik Schierl**.

Kathryn Hire, an astronaut and Navy reserve component Sailor assigned to the Office of Naval Research, presents items she took to space to Rear Adm. **Nevin Carr**.

Figure 6. **Box–name correspondences predicted by our model.** We show predicted entities on top of the their associated box (in white). Ground truth links are denoted by matching colors. Please refer to the supplemental material for additional qualitative results.

6.1. Comparison to Prior Work

We evaluate several recent visual grounding models on the *Who's Waldo* test set: a weakly-supervised framework by Gupta *et al.* [24], a supervised neural chain conditional random field that captures entity dependencies (SL-CCRF) [38], and a supervised network that combines attention from separate modules (MAttNet) [67]. We also evaluate UNITER [10], a pretrained multi-task vision-and-language framework, which our model is based on.

Table 1 shows test set accuracies for our approach and for existing methods trained on different datasets. We report 95% binomial proportion confidence intervals (Wilson score intervals) with these accuracies. For existing models, we vary how names are provided during inference because these models are not automatically compatible with our placeholder `[NAME]` token: (a) unmodified **full names**, (b) **random** popular names, or (c) a **constant** "person" string—*e.g.*, "Harry met Sally" is modified to "person met person".

We also evaluate several heuristics that illustrate the challenges and biases in our data (Table 1), such as a potential left-to-right bias for named individuals. In particular, we order the names in the caption from left to right, and pair them with detections sorted by (a) decreasing area (Big→Small), (b) left-to-right upper-left coordinates (L→R (All)), or (c) left-to-right upper-left coordinates with only the largest d detections (L→R (Largest)). We set $d = max(m, n)$ for m detections and n names. We also compare to random guessing. We observe that these heuristics yield non-trivial, and even strong, performances. This could be because realistic captions tend to follow a left-to-right ordering (especially for posed people—but see Figure 6 for counterexamples) and filtering by detection size can remove unreferred people. However, even the strongest heuristic leaves much room for improvement. These heuristics are also useful to frame the performance of pretrained visual grounding models. Supervised models (SL-CCRF and MAttNet) perform similarly to Big→Small, illustrating that these models may be utilizing size-related cues—especially MAttNet, which *only* processes names and not full sentences. We show qualitative results for all baselines in the supplemental material.

Method	Accuracy
Input features	
w/o visual features	55.4 ± 1.07
w/o spatial features	58.0 ± 1.06
w/o textual features	51.3 ± 1.07
spatial features only	31.2 ± 0.99
Learning	
w/o $\mathcal{L}_{\text{intra}}$	31.4 ± 1.00
w/o $\mathcal{L}_{\text{inter}}$	61.9 ± 1.04
w/o $\mathcal{L}_{\varnothing}$	61.7 ± 1.04
w/o pretraining	50.2 ± 1.07

Table 2. Ablation study, evaluating the effect of using different input features, loss terms and the impact using a pretrained model.

6.2. Ablation study

Table 2 shows ablation results. We train models using only a subset of features by ablating (i) visual features: set instead to a fixed representation, averaged over 1000 random detections; (ii) spatial features: fixed at image center coordinates; (iii) textual embeddings: with all words masked out, retaining only position features and special `[NAME]` tokens; and (iv) textual and visual embeddings: retaining only spatial features. The impact of each input modality is significant, with performance dropping by 5.5% for (ii) and 12.2% for (iii). While these ablations significantly limit the information available for this task, our model performs much better than random guessing in all cases, suggesting that it learns some data biases. Both (i) and (iii) are capable of learning a left-to-right association. Indeed, their correct matches significantly overlap with those of the "L→R (Largest)" heuristics, by 81.7% for (i) and 82.4% for (iii). Finally, from (iv) we infer that spatial features alone are not enough for learning such similarities.

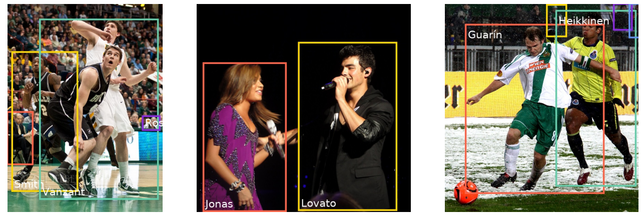
We also quantify the importance of each proposed objective. Training without estimated correspondences (*i.e.*, $\mathcal{L}_{\text{intra}}$) yields the most significant drop in performance, resulting in nearly random guessing. This illustrates the importance of supervised data for our task. Ablating the other losses ($\mathcal{L}_{\text{inter}}$ and $\mathcal{L}_{\varnothing}$) degrades performance by only 1.7%. The relatively small impact of $\mathcal{L}_{\text{inter}}$ highlights the importance of having many samples that capture interactions between multiple people, rather than samples with just one detection and referred person.

We also report the performance obtained by training our full model from scratch, without UNITER's pretrained weights. This leads to a large drop in performance ($> 13\%$).

6.3. Analysis of results

We analyze the performance of our model over different test subsets to better understand what the model is learning. We observe that our model is more robust to a larger number of faces compared to L→R (Largest). For instance, in the case of only one referred person in an image, our model retains high performance over an increasing number of faces, while the heuristic drops by almost 20% (from an accuracy of 84.5% for two detected people down to an accuracy of 67.6% for four or more detected people). We further demonstrate this breakdown in the supplemental material. Another subset we consider is an *interactive* subset of test samples (*i.e.* those with at least two detections and referred people and a verb in their caption). This potentially more challenging subset constitutes nearly one-third of our test set. Our model's performance drops to 52.1%, while the baseline performance drops to 45.0%.

We also analyzed whether having multiple mentions of a person's name affects performance. Approximately 3% of referred people in the test set are mentioned multiple times in the caption. For these identities, our model has a modest improvement of 2.1% if provided additional mentions. This illustrates that our model can leverage additional information from co-occurrences in a caption to some extent. Finally, we also analyze the performance of our method over several categories of identity occupations in the supplemental material, as we observe that these correlate well with different situations captured by our dataset.



Left: "Butler's **Andrew Smith** and Siena's **Ryan Rossiter** both try to anticipate the rebound, as Butler's **Shawn Vanzant** closes in from behind."
Center: " **Joe Jonas** and **Demi Lovato** performing in the Jonas Brothers Live In Concert."
Right:" **Markus Heikkinen** blocks **Freddy Guarín** ."

Figure 7. **Examples our model predicted incorrectly**, showing detected named entities in bold and entities linked with image regions in unique colors, corresponding to the boxes on the images.

6.4. Limitations

The complexity of certain interactions, such as in sports games where players compete closely together, poses challenges, not only to our model, but also to the person detector and our method for estimating ground truth links. Figure 7 (left) demonstrates an example of a basketball game where the bodies of players overlap, thus some are not detected. The example on the right illustrates a failure of our model, where the interaction "blocks" is not correctly interpreted.

Further, some captions are insufficient to produce meaningful links. For example, in Figure 7 (center), after replacing the names "Joe Jonas" and "Demi Lovato" with `[NAME]`, it is impossible to tell which performer each corresponds to. Hence our model resorts to a simple left-to-right heuristic.

7. Conclusion

We present a task, dataset, and method for linking people across images and text. By masking out names of people, we force methods to not memorize the appearance of specific individuals, but to understand contextual cues and interactions between multiple people. Our approach shows encouraging performance on this task, but also indicates that the underlying task is very challenging and, as such, there is ample room for improvement via future methods that leverage our data. In particular, the performance of all methods drops given examples involving actions (as indicated by captions with verbs) and as the number of people referred to in a caption grows, indicating unresolved challenges in scaling to complex scenarios.

Acknowledgments. This work was supported by the National Science Foundation (IIS-2008313, CAREER-1750499), a Google Focused Award, the generosity of Eric & Wendy Schmidt by recommendation of the Schmidt Futures program, and the Zuckerman STEM Leadership Program.

References

[1] Hassan Akbari, Svebor Karaman, Surabhi Bhargava, Brian Chen, Carl Vondrick, and Shih-Fu Chang. Multi-level multi-modal common semantic space for image-phrase grounding. In *ICCV*, pages 12476–12486, 2019.

[2] A. Akbik, T. Bergmann, Duncan Blythe, K. Rasul, Stefan Schweter, and Roland Vollgraf. Flair: An easy-to-use framework for state-of-the-art nlp. In *NAACL-HLT*, 2019.

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

[4] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683, 2018.

[5] Steven Bird. Nltk: The natural language toolkit. *ArXiv*, cs.CL/0205028, 2006.

[6] Roberto Brunelli and Daniele Falavigna. Person identification using multiple cues. *IEEE transactions on pattern analysis and machine intelligence*, 17(10):955–966, 1995.

[7] Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Conference on Computer Vision and Pattern Recognition*, 2019.

[8] Kan Chen, Jiyang Gao, and Ram Nevatia. Knowledge aided consistency for weakly supervised phrase grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4042–4050, 2018.

[9] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[10] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: Universal image-text representation learning. In *ECCV*, 2020.

[11] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. `https://github.com/open-mmlab/mmpose`, 2020.

[12] Kate Crawford and Trevor Paglen. Excavating ai: the politics of images in machine learning training sets. *Excavating AI*, 2019.

[13] Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2601–2610, 2019.

[14] Chaorui Deng, Qi Wu, Qingyao Wu, Fuyuan Hu, Fan Lyu, and Mingkui Tan. Visual grounding via accumulated attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7746–7755, 2018.

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009.

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

[17] Martin Engilberge, Louis Chevallier, Patrick Pérez, and Matthieu Cord. Finding beans in burgers: Deep semantic-visual embedding with localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3984–3993, 2018.

[18] Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *arXiv preprint arXiv:1909.00512*, 2019.

[19] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*, 2016.

[20] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 457–468, 2016.

[21] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *arXiv preprint arXiv:2006.06195*, 2020.

[22] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew E. Peters, M. Schmitz, and Luke Zettlemoyer. Allennlp: A deep semantic natural language processing platform. *ArXiv*, abs/1803.07640, 2018.

[23] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. Datasheets for datasets, 2018.

[24] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. *arXiv preprint arXiv:2006.09920*, 2020.

[25] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask r-cnn. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:386–397, 2020.

[26] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *The IEEE International Conference on Computer Vision*, pages 804–813, 2017.

[27] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[28] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in pho-

tographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.
[29] T. Kiss and J. Strunk. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32:485–525, 2006.
[30] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.
[31] H. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955.
[32] Ryan Layne, Timothy M Hospedales, and Shaogang Gong. Towards person identification and re-identification with attributes. In *European Conference on Computer Vision*, pages 402–412. Springer, 2012.
[33] Kenton Lee, Luheng He, M. Lewis, and Luke Zettlemoyer. End-to-end neural coreference resolution. *ArXiv*, abs/1707.07045, 2017.
[34] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.
[35] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.
[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.
[37] Daqing Liu, Hanwang Zhang, Feng Wu, and Zheng-Jun Zha. Learning to assemble neural module tree networks for visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4673–4682, 2019.
[38] Jiacheng Liu and Julia Hockenmaier. Phrase grounding by soft-label chain conditional random field. In *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 5112–5122. Association for Computational Linguistics, 2020.
[39] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019.
[40] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446, 2020.
[41] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, 2017.
[42] Junhua Mao, J. Huang, A. Toshev, Oana-Maria Camburu, A. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11–20, 2016.
[43] Dipendra Misra, John Langford, and Yoav Artzi. Mapping instructions and visual observations to actions with reinforcement learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1004–1015, 2017.
[44] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24:1143–1151, 2011.
[45] Said Pertuz, Domenec Puig, and Miguel Angel Garcia. Analysis of focus measure operators for shape-from-focus. *Pattern Recognition*, 46(5):1415–1432, 2013.
[46] Bryan A Plummer, Arun Mallya, Christopher M Cervantes, Julia Hockenmaier, and Svetlana Lazebnik. Phrase localization and visual relationship detection with comprehensive image-language cues. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1928–1937, 2017.
[47] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, pages 2641–2649, 2015.
[48] Siyuan Qiao, Liang-Chieh Chen, and Alan Yuille. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. *arXiv preprint arXiv:2006.02334*, 2020.
[49] Yanyuan Qiao, Chaorui Deng, and Qi Wu. Referring expression comprehension: A survey of methods and datasets. *IEEE Transactions on Multimedia*, 2020.
[50] Shaoqing Ren, Kaiming He, Ross B. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015.
[51] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pages 817–834. Springer, 2016.
[52] Arka Sadhu, Kan Chen, and Ram Nevatia. Zero-shot grounding of objects from natural language queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4694–4703, 2019.
[53] Amir Sadovnik, Andrew Gallagher, and Tsuhan Chen. It's not polite to point: Describing people with uncertain attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3089–3096, 2013.
[54] D. Sandberg. Face recognition using tensorflow. `https://github.com/davidsandberg/facenet`, 2017.
[55] Florian Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[56] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

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

[58] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *ICCV*, pages 7464–7473, 2019.

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

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[61] Mingzhe Wang, Mahmoud Azab, Noriyuki Kojima, Rada Mihalcea, and Jia Deng. Structured matching for phrase localization. In *European Conference on Computer Vision*, pages 696–711. Springer, 2016.

[62] Qinxin Wang, Hao Tan, Sheng Shen, Michael W Mahoney, and Zhewei Yao. Maf: Multimodal alignment framework for weakly-supervised phrase grounding. *arXiv preprint arXiv:2010.05379*, 2020.

[63] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[64] Fanyi Xiao, Leonid Sigal, and Yong Jae Lee. Weakly-supervised visual grounding of phrases with linguistic structures. In *CVPR*, pages 5945–5954, 2017.

[65] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 547–558, 2020.

[66] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *CVPR*, 2016.

[67] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315, 2018.

[68] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016.

[69] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7093–7102, 2020.

[70] Fang Zhao, Jianshu Li, Jian Zhao, and Jiashi Feng. Weakly supervised phrase localization with multi-scale anchored transformer network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5696–5705, 2018.

[71] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.

[72] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proc. AAAI Conference on Artificial Intelligence*, pages 13041–13049, 2020.