Who's Waldo? Linking People Across Text and Images — Supplementary Material (ICCV 2021) —

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

https://whoswaldo.github.io

Contents

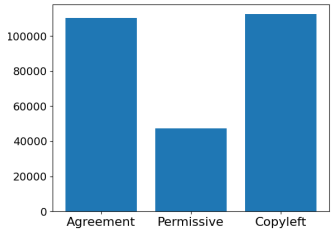


Figure 1. Distribution of copyright license groups for images in *Who's Waldo*.

A. Dataset Visualizations and Details

Please refer to the following URL for samples from our *Who's Waldo* dataset: https://whoswaldo.github.io/dataset_examples.html

Our dataset has 215K ground truth links in total (for 193K images). Our dataset originates from over 400K Wikimedia identities and has ground truth links for 93K.

All images originate from Wikimedia Commons under free licenses. We group the licenses by freedom[1] as in Table 1.

We include a word cloud of the verbs present in our dataset in Figure 2.

[1]https://en.wikipedia.org/wiki/Free_license#By_freedom

Agreement	Copyrighted free use, No restrictions, CC0 Public Domain, Public domain, WTFPL
Permissive	MIT, BSD, CC BY 1.0, CC BY 2.0, CC BY 2.5, CC BY 3.0, CC BY 4.0, Attribution, OGDL, Licence Ouverte, KOGL Type 1, OGL-C 2.0, OSPL, GODL-India, Beerware
Copyleft	GPL, GPLv2, GPLv3, LGPL, CC SA 1.0, CC BY-SA 2.0, CC BY-SA 2.5, CC BY-SA 3.0, CC BY-SA 4.0, Nagi BY SA, GFDL 1.1, GFDL 1.2, GFDL 1.3, GFDL, ODbL, OGL, OGL 2, OGL 3, FAL, CeCILL

Table 1. Free licenses for images in our dataset (organized by freedom).

Figure 2. Visualization of verbs appearing in our dataset's captions. Larger font size correspond to verbs that appear more frequently in the dataset.

Our dataset contains images for at least 263K male and 70K female Wikimedia identities (these are identities we have labels for). We acknowledge this imbalance in ratio and attribute this to existing biases in our data source. However, our dataset is large enough that one could sample a more balanced subset. Our dataset does present diversity in the occupations of identities, as can be seen in Figure 3.

We show distributions of image resolutions in Figure 5

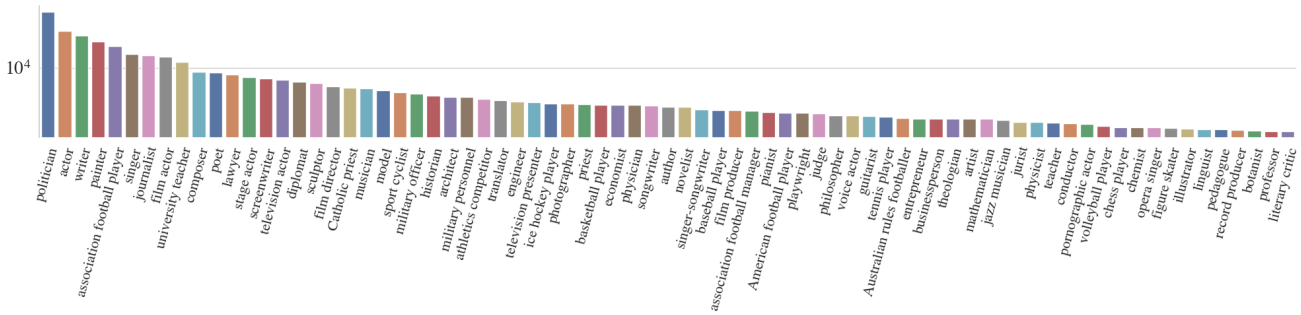


Figure 3. Distribution of occupations for Wikimedia identities in our dataset.

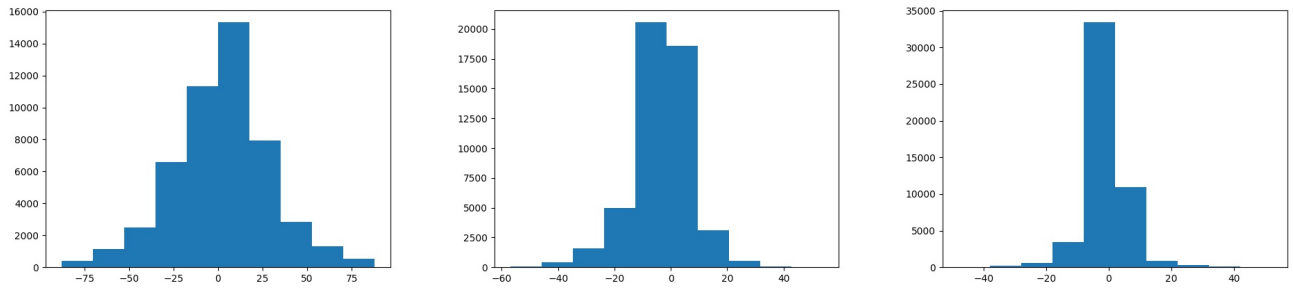


Figure 4. Distribution of faces by degree of pose (head orientation) from a random subset of 50,000 detections. Note: yaw is the primary indicator of diversity in pose, as pitch and roll are limited by physical constraints for head rotation.

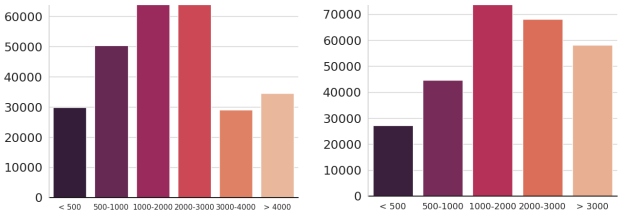


Figure 5. Distribution of image resolutions in our dataset.

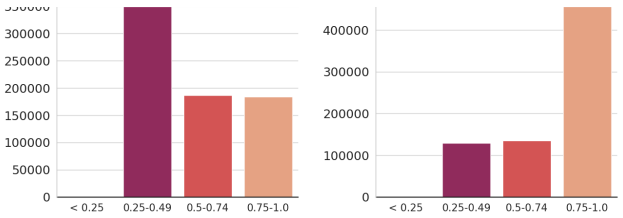


Figure 6. Distribution of detection sizes in our dataset.

and sizes of detection boxes (relative to image sizes) in Figure 6. We show a distribution of head poses in our dataset in Figure 4.

B. Implementation Details

B.1. Dataset Cleaning

We filter our data by removing all examples in which there are no people detected in an image or no people referred to in captions. We remove examples with captions that don't contain verbs or words other than names and stop words (*i.e.* insubstantial captions). We further cleanse this data by removing images taken before 1990 (according to metadata) as we found this was a significant source of noise. We also found the presence of "cropped" versions of images that can be detected directly from file names containing the word "cropped", which usually only picture one person but have captions implying the presence of multiple, and also removed these.

B.2. Training details

We download the pretrained UNITER [2] model (UNITER-base). We use the "bert-base-cased" vocabulary from pytorch-transformers and add the [NAME] token. Following their implementation[2], we define two training tasks that use two non-overlapping subsets of our dataset: $(1, 1)$, containing images with exactly one referred person and one person box detected in the image, and (m, n), containing

[2] https://github.com/ChenRocks/UNITER

Figure 7. For each referred person associated with a "primary" image on Wikimedia Commons (right), we compute face dissimilarities between the face in the "primary" image and all detected faces. By finding a minimum weight bipartite matching (over all referred people), we recover a partial matching from referred people to detections (for simplicity, we only show these dissimilarities for a single referred person and for a subset of faces in the image). The estimated link is shown in blue.

all other images (*i.e*. more than one referred person *or* more than one box).

The first task, denoted as **Task-1-1**, trains on the $(1, 1)$ subset using the $\mathcal{L}_{\text{inter}}$ objective, with 0.5 probability of negative sampled image-caption pairs. The second task, denoted as **Task-M-N**, trains on the (m, n) subset using the $\mathcal{L}_{\text{intra}}$ and $\mathcal{L}_{\varnothing}$ objectives. Furthermore, regarding $\mathcal{L}_{\text{intra}}$, we note that this loss a sum over two cross-entropy losses, one over different boxes in the image and the other over different names in the caption. Task-1-1 and Task-M-N are trained using a $1 : 2$ ratio.

We train 50,000 steps, validating performance over the validation set every 500 steps, with batch size of 1024. The max caption length we consider is 60 tokens, and the number of bounding boxes we consider is between 1 and 100, inclusive. Image-caption pairs not within these boundaries are filtered out during training. We use a learning rate of $5e - 5$, weight decay of 0.01 and dropout 0.1, consistent with the default UNITER parameters (all other parameters are also set according to their default values).

C. Baselines

Next we provide more details on how we obtain the reported scores on the pretrained models we evaluate on the WikiPeople test set.

Gupta *et al*. [4]. We download their two pretrained models, trained on either COCO [7] or Flickr30 Entities [11], from their official code repository[3]. Following their implementation, visual features are extracted using the Bottom-Up Attention model [1] yielding a 2048-d visual representation. A pretrained BERT [3] model is used to extract 768-d contextualized word representations. We follow their evaluation protocol and compute a phrase-level attention score for each box by taking the maximum attention score assigned to the box by any of the tokens in the name. The boxes are then ranked according to this phrase level score, with the maximum scoring box selected as the corresponding box. This top-scoring box is compared with the ground-truth box.

[3] https://github.com/BigRedT/info-ground

Method	Accuracy
Gupta et al. [4]	31.78
SL-CCRF [9]	30.07
MAttNet [13]	27.53

Table 2. Performance obtained on the baselines trained on our data. As further detailed in the text, these baselines cannot be naively adapted for our task.

SL-CCRF [8]. We download the pretrained "Soft-Label Chain CRF Model" from their official code repository[4], which yields the highest performance among their available models. Following their implementation, visual features are extracted using the Bottom-Up Attention model [1] yielding a 2048-d visual representation. We use their all default parameters, as follows: 1024-d contextualized word embeddings, the maximum number of mentions is set to 25, and a 5-d spatial feature is concatenated with the visual features. The number of regions proposals are according to the number of detected people boxes. However, as their model also includes a regression bounding box loss, their final predictions aren't entirely aligned with the input bounding boxes. We account for that gap in the evaluation, by considering boxes with IoU ≥ 0.5.

MAttNet [13]. We downloaded a model from the official repository[5] that was pretrained on the RefCOCOg dataset [10]. Following their implementation, visual features are extracted using a modified implementation of Mask R-CNN [5], as specified by the authors [13]. However, we provide our own bounding boxes and compute Faster R-CNN region features [12] over these, instead of using their proposals. A Language Attention Network with bidirectional LSTMs (as specified by MAttNet [13]) is used to extract phrase embeddings. We use these modules to predict a detection for each individual referring expression (*i.e*. a person's name).

D. Additional Results and Ablations

We report performance obtained on all three baselines while training on our data in Table 2. The low performance obtained on the baselines is not surprising as (1) weakly supervised techniques (such as Gupta et al. [4]) do not have access to ground truth supervision—in our ablations this similarly results in a significant performance drop; (2) phrase grounding techniques (such as MAttNet [13]) only

[4] https://github.com/liujch1998/SoftLabelCCRF
[5] https://github.com/lichengunc/MAttNet

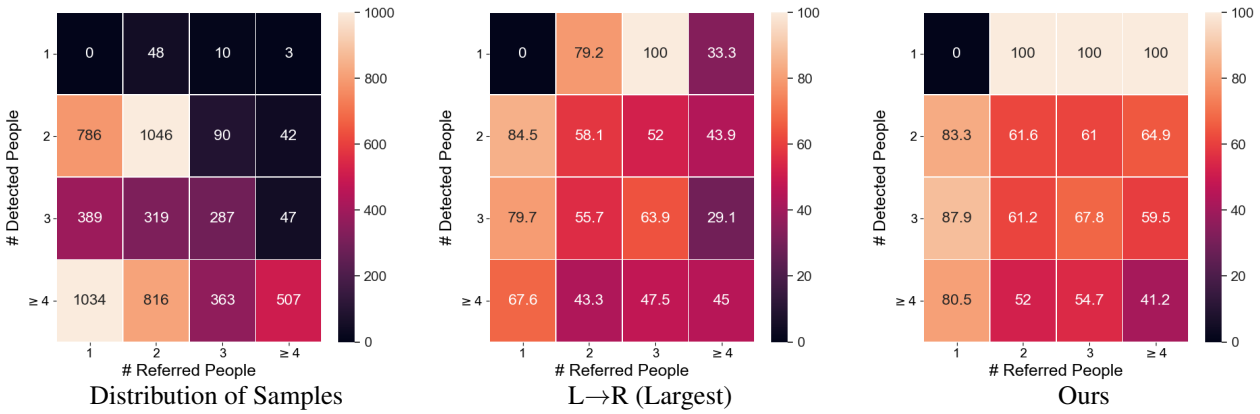


Figure 8. Accuracy breakdown by number of referred people and detected faces for our model and the L→R (Largest) baseline. The test set sample distribution is illustrated on the left (no images with just a single detection and referred person are included in our evaluation).

Method	Max-Box	Bipartite
Input features		
w/o visual features	55.4	54.2
w/o spatial features	58.0	57.0
w/o textual features	51.3	50.8
Learning		
w/o $\mathcal{L}_{\text{intra}}$	31.4	30.1
w/o $\mathcal{L}_{\text{inter}}$	61.9	60.9
w/o $\mathcal{L}_{\varnothing}$	61.7	61.2
w/o pretraining	50.2	38.7
w/ optimal transport loss	62.2	61.6
Ours (full)	63.5	61.9

Table 3. Ablation study, evaluating the effect of using a bipartite matching algorithm during inference (second column) and using an additional optimal transport loss (second to last row).

process the phrase describing the region (which would be masked out in our case); and (3) SL-CCRF also processes the masked out phrases, along with dependencies between string-adjacent phrases (which evidently are not enough on their own for the model to learn meaningful grounding).

All results reported in the paper are obtained by selecting, for each referred person, the most similar box according to S. In Table 3, we also report performance by performing a minimum weight bipartite matching [6] over the similarity matrix, thus producing a natural one-to-one mapping. As illustrated in the table, this yields a decrease in performance of approximately 1%. We also train a model with an additional (unsupervised) optimal transport loss, which was proposed for pretraining the UNITER [2] model, as it encourages sparsity, and could potentially improve alignments between words and regions in the image (or names and people's boxes in our case). Results show that adding this loss on top of S does not yield an improvement in performance (and even slightly degrades our full model's performance). This suggests that robust alignments are achieved from the training supervision directly, without need for additional regularization.

Figure 8 illustrates the distribution of samples and performance breakdowns for L→R (Largest) and our model over the numbers of referred people in a caption (n) and people detected in an image (m). We compute average accuracies over all relevant test subset images. As illustrated in the figure, the heuristic surpasses our model over only two subsets—($m = 2, n = 1$) and ($m \geq 4, n \geq 4$), given m detections and n referred people—and performs worse in all other subsets.

We find that occupations correlate with different situations—images featuring athletes, for instance, have different properties from those featuring singers. We observe that model performance varies somewhat across different occupation types. For instance, considering only the interactive subset of test samples, accuracy on people with athletic occupations (association football player, basketball player, etc.) is lower than accuracy for politicians or performers (actor, model, musician, etc.), while their distribution in the training set is similar (athletes, politicians, and performers are each captured by 10–13$\%$ of the interactive training set). A potential explanation is that interactions within sports-themed images are broader and more complex than in other categories.

We also observe that over the full set test, performance over politician samples is significantly lower, and this is also reflected in a lower left-to-right ordering accuracy. A visual analysis reveals that these samples are indeed more challenging, as in many cases the captions mostly mention notable individuals regardless of the visual arrangement of the captured individuals.

Finally, we experiment with training models using several forms of standard augmentation techniques. Results are reported in Table 5. Note that the nature of our dataset and task renders some augmentations more sensible than others.

Set	Politicians	Athletes	Performers
Interactive			
L→R (Largest)	47.1	43.0	49.8
Ours	52.5	51.1	54.9
All			
L→R (Largest)	52.4	70.6	67.4
Ours	54.8	76.3	71.2

Table 4. Analyzing model performance by identity occupation for the interactive subset and for all data samples. Test accuracy for the strongest baseline and for our model is reported for samples belonging to the occupation categories specified on top.

Augmentation	Accuracy
Ours	63.5
w/ horizontal flips	53.8
w/ translations	62.0
w/ color jittering	63.0

Table 5. Evaluating the effect of using standard data augmentation techniques during training.

In particular, a model trained with random horizontal flipping yields significantly lower performance. This is likely due to the inherent left-to-right ordering in the images and captions, as some captions in our dataset either explicit annotate people with "(left)" and "(right)", or implicitly mention people in the left-to-right order they appear in the image. Other augmentations, such as translating all bounding boxes within the image or performing random color jittering on the images, yields comparable performance.

E. Additional Qualitative Results

Figure 9 shows additional visualizations of our model's predictions for samples in our test set. Figure 10 and Figure 11 respectively show results obtained with prior supervised and weakly-supervised grounding models. As illustrated in the figures, prior visual grounding works struggle in correctly linking people across images and text for these challenging examples, which cover various interactions between multiple people. Errors can be attributed to selecting a single box for all referred people, or selecting (smaller) boxes that are unreferenced to in the caption.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] H. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955.

[7] 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.

[8] 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.

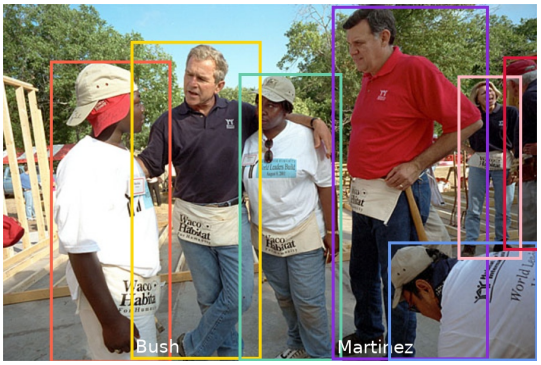
[9] Jiacheng Liu and Julia Hockenmaier. Phrase grounding by soft-label chain conditional random field. In *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 5112–5122. Association for Computational Linguistics, 2020.

[10] 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.

[11] 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.

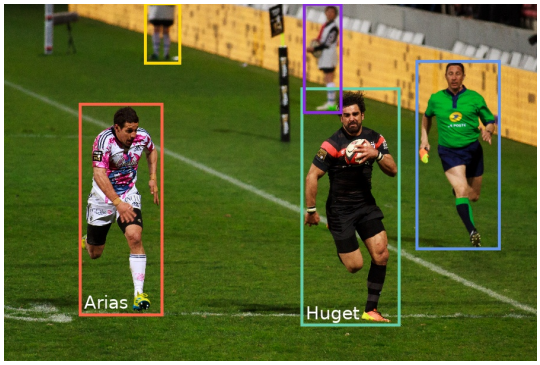
[12] Shaoqing Ren, Kaiming He, Ross B. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015.

[13] 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.

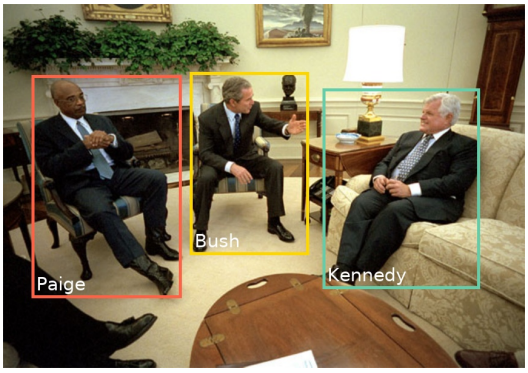


President **Bush** and Secretary for Housing and Urban Development **Martinez**, far right, talk with new friends during a break from their house-building efforts at the Waco, Texas, location of Habitat for Humanity's "World Leaders Build" construction drive August 8, 2001.

Mohamed Bamba dunks in front of **Collin Sexton** at the McDonald's All-American Boys Game.

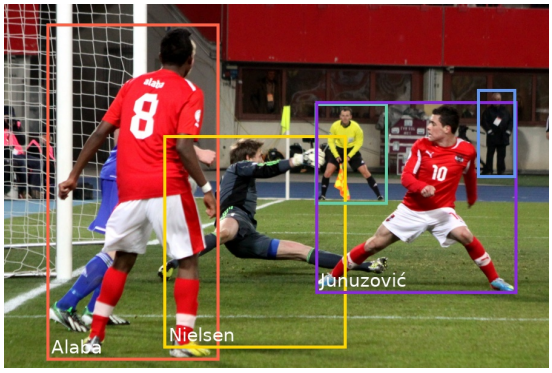


Yoann Huget out run **Julien Arias** to score his second try of the match during Stade toulousain vs Stade français Paris, March 24th, 2012.

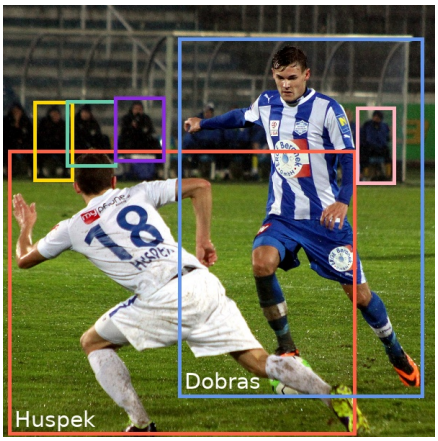


President **Bush** meets with Secretary of Education **Rod Paige**, left, and Senator **Edward Kennedy** August 2, 2001, to discuss the education reforms for the country.

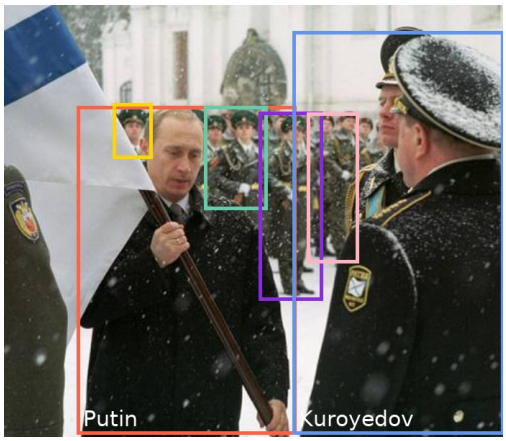
Markelle Fultz shoots over **Kyle Guy** at the McDonald's All-American Boys Game.



The photo shows **David Alaba** (Austria), **Gunnar Nielsen** (Faroe Islands) **Zlatko Junuzović** (Austria).



The photo shows **Kristijan Dobras** (SC Wiener Neustadt, blue shirt) and Philipp Huspek (SV Grödig, white shirt).



President **Putin** presenting the banner of the Navy to its Commander-in-Chief Admiral Vladimir Kuroyedov.

Astronaut **Terrence W. Wilcutt**, STS-68 pilot, goes over his notes. Checking the notes is **Alan M. Rochford**, suit expert.

Figure 9. **Additional 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.

Commandant of the U.S. Marine Corps Gen. **James F. Amos**, left, participates in a gift exchange with Commandant General of the British Royal Marines Maj. Gen. **Ed Davis**.

Caleb Marchbank kicking away from **Matt de Boer** during the AFL round twelve match between Carlton and Greater Western Sydney on 11 June 2017 at Etihad Stadium.

Justise Winslow of the Miami Heat defending **LeBron James**.

Figure 10. Comparing against supervised visual grounding techniques, SL-CCRF [9] and MAttNet [13], and the pretrained UNITER [2] model. We show predicted entities on top of the their associated box (in white). Ground truth links are denoted by matching colors. For SL-CCRF [9], as their model incorporates a regression loss that modifies the input boxes, we only show the predicted boxes. In both SL-CCRF [9] and MAttNet [13], errors are attributed to selecting the same box for multiple referred people. It should be noted that this is not always the case, and from further visual inspection, in many cases these models are capable of selecting multiple boxes. We can see that the pretrained UNITER model provides unique assignments for all three examples, possibly due to the optimal transport loss they propose to encourage robust word-region alignments. The selected boxes, however, are only accurate in the middle example (and partially accurate in the leftmost example).

Commandant of the U.S. Marine Corps Gen. **James F. Amos**, left, participates in a gift exchange with Commandant General of the British Royal Marines Maj. Gen. **Ed Davis**.

Caleb Marchbank kicking away from **Matt de Boer** during the AFL round twelve match between Carlton and Greater Western Sydney on 11 June 2017 at Etihad Stadium.

Justise Winslow of the Miami Heat defending **LeBron James**.

Figure 11. Comparing against the weakly-supervised visual grounding technique proposed by Gupta *et al*. [4]. We evaluate on both of their pretrained models, trained on Flickr30K Entities [11] (top row) and COCO [7] (second row). We show predicted entities on top of the their associated box (in white). Ground truth links are denoted by matching colors. Errors are attributed to either selecting the same box for multiple referred people (*e.g*. rightmost example), or selecting irrelevant boxes, such as the yellow box in the middle image, top row, or the orange box in the left image, second row.