Phrase Localization Without Paired Training Examples

Josiah Wang
Imperial College London
http://www.josiahwang.com

Lucia Specia
Imperial College London
l.specia@imperial.ac.uk

Abstract

Localizing phrases in images is an important part of image understanding and can be useful in many applications that require mappings between textual and visual information. Existing work attempts to learn these mappings from examples of phrase-image region correspondences (strong supervision) or from phrase-image pairs (weak supervision). We postulate that such paired annotations are unnecessary, and propose the first method for the phrase localization problem where neither training procedure nor paired, task-specific data is required. Our method is simple but effective: we use off-the-shelf approaches to detect objects, scenes and colours in images, and explore different approaches to measure semantic similarity between the categories of detected visual elements and words in phrases. Experiments on two well-known phrase localization datasets show that this approach surpasses all weakly supervised methods by a large margin and performs very competitively to strongly supervised methods, and can thus be considered a strong baseline to the task. The non-paired nature of our method makes it applicable to any domain and where no paired phrase localization annotation is available.

1. Introduction

Significant progress has been made in recent years in the task of detecting and localizing instances of object categories in images, especially with deep convolutional neural network (CNN) approaches to object detection [7, 8, 10, 19, 26, 27, 28, 31]. In most work, object detection labels are treated as a fixed set of category labels, and visual detectors are trained to localize each category in the image. In more realistic applications, however, people refer to objects in images via free-form textual phrases, instead of object categories. For example, *a brown and furry puppy* instead of *dog*. Phrase-level localization has been introduced [11, 15, 17, 20, 24, 37] to address this need by combining visual object recognition and natural language processing. What we refer to as 'phrases' can include single words, short clauses or phrases, or even complete sentences. All previous work assume some form of supervision at training time: either strong supervision (object localization for the phrase in the image is provided) [2, 3, 11, 12, 22, 23, 29, 33, 39] or weak supervision (the phrase and image pair is provided, but not the object's localization in the image) [1, 34, 35, 40] (Figure 1). Such specific bounding box annotations and even image-phrase pairs, however, are hard and labourious to obtain. This makes it difficult to scale detection up to more realistic settings covering the large space of possible phrases that can be uttered by a person.

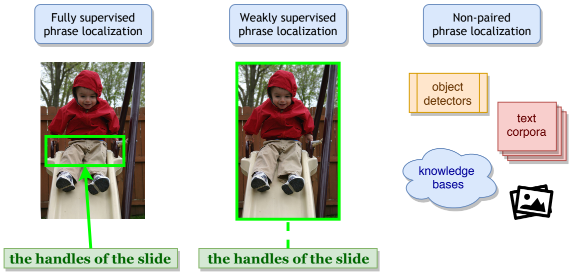
Fully supervised phrase localization
Weakly supervised phrase localization
Non-paired phrase localization
object detectors
text corpora
knowledge bases
the handles of the slide
the handles of the slide

Figure 1. We investigate the task of phrase localization *without* paired training examples. Conventional settings require phrase and image localization annotations (fully supervised) or phrase and image pairs (weakly supervised) at training time. In contrast, the non-paired setting does not provide such annotations for training, but instead allows models to exploit resources such as off-the-shelf visual detectors, large-scale general corpora, knowledge bases and generic images to localize previously unseen phrases at test time. This non-paired setting is thus a baseline to supervised settings.

In this paper, we tackle the novel task of **phrase localization in images *without* any paired examples**, *i.e*. the model has access to neither phrase-image pairs nor their localization in the image at training time (a 'training' phase may not even be required). To our knowledge, no previous

* The authors thank the anonymous reviewers and area chairs for their feedback. This work was supported by the MultiMT project (H2020 ERC Starting Grant No. 678017); and by the MMVC project, via an Institutional Links grant (ID 352343575) under the Newton-Katip Çelebi Fund partnership funded by the UK Department of Business, Energy and Industrial Strategy (BEIS) and Scientific and Technological Research Council of Turkey (TÜBİTAK) and delivered by the British Council.

work has explored this challenging setting of performing phrase localization without paired annotations (image-level or object-level). We argue that such a 'non-paired' setting better reflects how humans localize objects in images – not by memorizing paired examples, but by assembling prior knowledge from more general sources and tasks (*e.g.* recognizing concepts or attributes) to tackle a more specialized task (phrase localization). Thus, this setting acts as a strong baseline to the phrase localization task, *i.e.* it demonstrates the extent to which a system can perform phrase localization even without having seen any such examples. This can give further insights into how paired examples can be better utilized for phrase localization in an informed manner, on top of what can be done without paired examples. The approach is also scalable to any domain and to any number of natural language and image pairs.

The main contribution of this paper is a model for phrase localization that is *not* trained on phrase localization annotations (Section 3). Instead, it exploits readily available resources, tools and external knowledge. Our model has the advantage of being **simple** and **interpretable**, acting as a strong baseline for the novel, non-paired setting. We provide an in-depth analysis of this model on two existing phrase localization datasets (Section 4), using different detectors and combination of detectors, semantic similarity measures for concept selection, and strategies to combine these components to localize previously unseen phrases.

Our experiments on two existing phrase localization datasets show that our approach without paired examples outperforms state-of-the-art weakly supervised models by a large margin, and is on par with fully supervised approaches that utilize large sets of annotated phrase localization examples and domain-specific tools at training time. The results suggest that, for these datasets, training with phrase localization annotations may not be necessary or optimal for tackling the phrase localization task.

2. Related work

The availability of datasets annotated with bounding box labels [4, 30] has allowed the development of deep CNN based detectors [7, 8, 10, 19, 26, 27, 28, 31], propelling the field of object recognition to produce more accurate detections and localization of object instances in images.

There has been recent interest in localizing objects using free-form natural language phrases instead of fixed labels, with datasets constructed for such tasks [15, 17, 20, 24, 37]. We classify existing phrase localization or grounding approaches as either strongly/fully supervised [2, 3, 11, 12, 22, 23, 29, 33, 39] or weakly supervised [1, 14, 34, 35, 40].

Methods that use strong supervision include those that project phrases and image regions onto a common space [23, 25, 33], those that build a language model for the phrase conditioned on bounding box proposals [12], and those that learn to attend to the correct region proposals given the phrase [29]. More recent approaches include conditioning an object detector on phrases instead of fixed object labels [11], leveraging semantic context and learning to regress bounding boxes directly from phrase localization data instead of relying on external region proposals [2, 3], and conditioning embeddings on the categories/groups to which a phrase belongs [22].

In a weakly supervised setting, no localization is provided at training time. Thus, such methods use external region proposals [1, 40], generic object category detectors [35] and attention maps [34] for localization. To learn to associate these region proposals with phrases, various methods have been proposed, including learning to constrain the regions' spatial positions using the parse tree of the caption [34], performing a continuous search using region proposals as anchors [40], linking words in text and detection labels using co-occurrence statistics from paired captions [35], and enforcing consistency between the concept labels of a region proposal and words in the query [1].

We are unaware of work that tackles phrase localization without paired examples. Yeh *et al.* [35] define their work as 'unsupervised', but we consider it weakly supervised. Their model uses image-phrase pairs from the training dataset (similar distribution as the test set) to compute co-occurrence statistics between words and concepts, and to train image classifiers for words in the phrases. Our model adapts Yeh *et al.*'s approach for when no paired training examples are available. In addition, we propose a new localization module that makes better direct use of the output of multiple detectors for phrase localization.

3. Model for non-paired phrase localization

Task definition. Given an image I and a query phrase q at *test* time, the aim of the phrase localization task is to produce a bounding box b encompassing the visual entity in I to which q refers. In contrast to conventional supervised settings, in our proposed non-paired setting annotated paired training examples (q,I) or (q,I,b) are *not* available at training or model construction time. Instead, models are allowed to use external resources that are not specific to phrase localization, for example general visual object detectors, generic text corpora, knowledge bases and thesauri, and images from generic datasets not annotated with phrases. We note that visual detectors may be trained in a supervised manner (*e.g.* with COCO or ImageNet), but there is no supervision in terms of phrase-based labels for the phrase localization task. Similarly, language models trained from generic text corpora may contain phrases from the test set, as long as they are independent of the images.

Our model builds upon the approach of Yeh *et al.* [35]. In contrast to their approach, however, we perform phrase localization without an explicit training step or phrase lo-

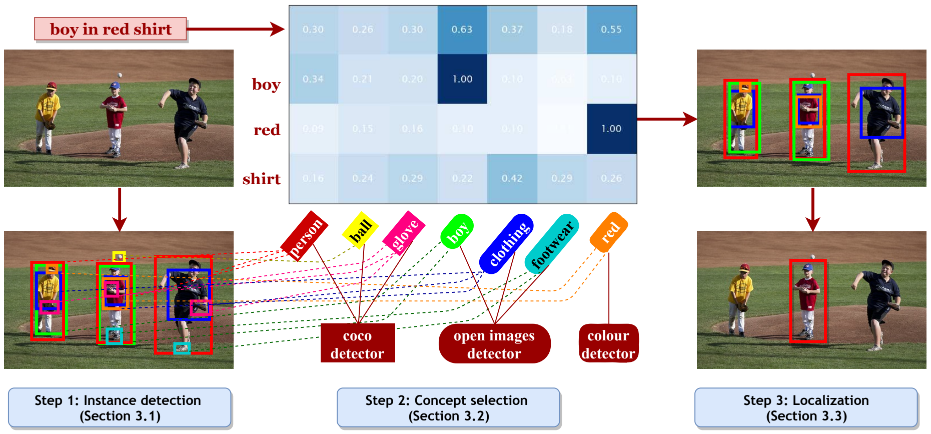

Figure 2. The three stages of the proposed model for non-paired phrase localization. The **instance detection** phase detects instances of various concepts using pre-trained detectors. The **concept selection** stage ranks these detected concepts against the query phrase (using pre-trained word embeddings) and forwards the best candidate concept instance(s) to the **localization** phase, where the model predicts the final bounding box for the query phrase.

calization annotations. We (i) incorporate a semantic similarity measure derived from general text corpora rather than aligned training examples; (ii) explore an array of off-the-shelf visual detectors not specifically trained for phrase localization; (iii) propose different strategies to perform phrase localization from detection outputs, including a novel consensus-based method that combines the output of multiple detectors.

At test time, our model performs phrase localization using a three-step process (Figure 2). In the first step – **instance detection** – it predicts candidate bounding boxes using a combination of different visual detectors (Section 3.1). In the second step – **concept selection** – the model computes the semantic similarity between the query phrase and the concept labels for instances detected in the previous step, and selects the most relevant instance(s) (Section 3.2). In the final step – **localization** – the model predicts the bounding box for the query phrase from the selected candidate instance(s) from the second step (Section 3.3).

3.1. Instance detection

The first stage of our non-paired phrase localization model relies on different visual object detectors. We explore using the detectors in isolation and by combining their output, where the concepts are not necessarily mutually exclusive. The key idea is to exploit the redundancy from multiple detectors to handle missing detections and to increase the importance of object instances detected across multiple detector groups. We experiment with the following:

1. **tfcoco**: A Faster R-CNN [28] detector trained to detect the 80 categories of MS COCO [18], using the Tensorflow Object Detection API [13],[1] with confidence threshold of 0.1.
2. **tfcoco20**: A subset of **tfcoco**, where we only consider the subset of 20 categories from PASCAL VOC [5]. This enables comparison to previous work.
3. **tfoid**: Another Faster R-CNN detector, trained to detect the 545 object categories of the Open Images Dataset (V2) [16], again using the TensorFlow Object Detection API,[2] with confidence threshold of 0.1.
4. **places365**: A WideResNet18 [38] classifier trained on the Places2 dataset [41] for 365 scene categories. We assume that scenes usually cover the full image, and return the whole image as the bounding box localization when the classification confidence is at least 0.1. We keep only the top 20 predicted classes.
5. **yolo9000**: A YOLO9000 detector [27] trained on MS COCO and ILSVRC [30] for 9,413 categories in a weakly supervised fashion. We use YOLOv2.
6. **colour**: A colour detector for 11 basic English colour terms, derived from the posterior across the colour terms for RGB pixels as learned from real world images [32]. We performed connected component labelling (8-connectivity) after thresholding the posteriors at 0.3 and generated bounding boxes for each labelled connected component. The area of the bounding boxes is constrained to be at least 625 pixels.

The detectors vary in accuracy and the number and type of categories covered. It is worth noting that none of the detectors above directly use images or phrase localization annotations from our test datasets. This will emphasize the ability of our phrase localization model to generalize to unseen data. More detectors could potentially be used to further improve recall, but the ones used here are sufficient to show that the proposed approach is very promising.

[1] `faster_rcnn_inception_resnet_v2_atrous_coco`

[2] `faster_rcnn_inception_resnet_v2_atrous_oid`

3.2. Concept selection

The second stage of our model bridges the query phrase to be localized and the output of detectors from Section 3.1. It computes the semantic similarity between each phrase and the detector concept labels. The intuition is that the detected instance of a concept that is very similar or related to a word or phrase in the query is most likely to be the target object. For example, the word *dancer* might be highly similar or related to the category *person*; thus even without a *dancer* detector, the model can infer that the detected *person* is likely to be the *dancer* mentioned in the query.

We represent both queries q and concept labels c as 300-dimensional CBOW word2vec embeddings [21]. Multi-word phrases are represented by the sum of the word vectors of each in-vocabulary word of the phrase, normalized to unit vector by its L_2-*norm*.[3] All words in the queries and concept labels (except **yolo9000**) are lowercased. For **yolo9000**, each category is a WordNet [6] synset. Thus we represent each category as the sum of the word vectors for each term in its synset, normalized to unit vector. Out-of-vocabulary words are addressed by matching case variants of the words (*Scotch whiskey* to *scotch whiskey*). Failing that, we attempt to match multiword phrases like before.

We noticed many misspellings among the query phrases. Thus, the model exploits another external resource to perform automated spell correction[4] for out-of-vocabulary words. The model finds candidate replacement words from word2vec's vocabulary and choosing the one with the highest frequency in the corpus used to train the embeddings. The model consistently obtained slightly higher accuracies with spell correction, and thus we report only the results with the spell-corrected queries.

We explore two approaches to aggregate the words in query phrases: as a single vector by summing the word vectors and normalizing to unit vector (**w2v-avg**), or by representing each word individually (**w2v**) and using only one of the words for localization (see Section 3.3).

We use cosine similarity as the semantic similarity measure $S(q, c)$ between a query q and a concept label c. This stage outputs a ranked list of candidate bounding box detections based on their similarity to the query phrase.

3.3. Localization

In the final stage, our proposed model predicts a bounding box given the query phrase and the ranked list of candidate detections from Section 3.2. This is accomplished by selecting from or aggregating the candidate instances that are most semantically similar to the query.

The simplest localization approach is to select from candidate detections the concept most similar to the query phrase. Where multiple instances of the same concepts are detected, we experiment with different tie-breaking strategies: (i) selecting a **random** instance; (ii) selecting the instance with the **largest** bounding box; (iii) selecting the instance with the highest class prediction **confidence**; (iv) generating a minimal bounding box enclosing *all* instances (**union**). The latter may useful for dealing with queries referring to multiple instances of an object (*e.g.* localizing *three people* from three individual *person* detections).

Besides simple heuristics, we also propose a novel tie-breaking approach by **consensus**. The main idea is that detectors can vote on the most likely localization, exploiting the redundancy across detectors and the different aspects of the phrase (*blue shirt*). We consider the semantic similarity of instances from the top-K concepts above a similarity threshold (we use K=5 and threshold 0.6). For each concept c_i, a pixel-level heatmap for the image, $M_{c_i}(I)$ is generated by setting to 1 pixels that overlap with any bounding box instance of the concept, and setting to 0 those that do not. We generate a combined heatmap $\hat{M}(I)$ by summing the heatmaps for each concept, each weighted by the semantic similarity score $S(q, c)$ from Section 3.2:

$$\hat{M}(I) = \sum_{i=1}^{K} S(q, c_i) M_{c_i}(I) \tag{1}$$

Phrase localization is performed by selecting the bounding box instances that voted for the pixels with the highest values, and choosing the box with the highest semantic similarity score as the predicted localization. In cases where there are multiple top scoring boxes, the model predict a minimal bounding box that encloses all such boxes.

We compare using a single combined word embedding for the phrase (**w2v-avg**) or using the embedding for one word to represent the phrase (**w2v**). For the latter, we can select the word with the highest semantic similarity to any detected concepts (**w2v-max**). Intuitively, we only consider one word from the phrase for localization, where this word has the highest similarity to a detected concept. Alternatively, we can use the *last* word for localization (**w2v-last**), assuming the last word is the head word. We default to localizing to the whole image when no words in the phrase are found in the vocabulary.

4. Experimental results

We evaluate our proposed models on two challenging datasets: **Flickr30kEntities** (Section 4.2) and **ReferItGame** (Section 4.3). Both have been used to evaluate supervised phrase localization [1, 29, 35]. Each dataset represent different challenges: Flickr30kEntities are noun phrases extracted from full image captions, while ReferItGame are short phrases generated from an interactive game where one player tries to localize the object the

[3] Averaging word embeddings for the entire phrase leads to the same experimental results.

[4] https://pypi.org/project/pyspellchecker/

other player is describing. Thus, we consider the latter as more challenging. We also test selected models on Visual Genome (Section 4.4) to investigate the model's scalability to a different dataset with sentence-level descriptions.

4.1. Evaluation metric

As in previous work, we use the *accuracy* metric for evaluation[5], where a predicted bounding box p_i for a query phrase is considered correct if its intersection over union (IoU) with the ground truth g_i is at least 50%.

For reference, we measure the extent to which the correct localization can be found among the candidate localizations from the concept selection stage (Section 3.2), depending on the similarity measure and the detector used. This *upperbound* accuracy is computed across N test instances as

$$\frac{1}{N}\sum_{i=1}^{N}\max_{j=1}^{B}\mathbb{1}\big(IoU(g_i,b_j)\geq 0.5\big) \tag{2}$$

where $\mathbb{1}(\cdot)$ is the indicator function and B the number of candidate bounding boxes. We report a version of the upperbound that additionally includes the minimal bounding box that encompasses the *union* of all candidates (thus $B+1$ candidates). This variant consistently gave higher upperbound accuracies than without the union. The results of both variants are given in the supplementary document.

4.2. Phrase localization on Flickr30kEntities

Flickr30kEntities [24] is based on Flickr30k [36], containing bounding box annotations for noun phrases occurring in the corresponding image captions. The test split [25] comprises $14,481$ phrases for $1,000$ images, which we use for evaluation. The training and validation splits are not used in our non-paired localization experiments.

As no non-paired phrase localization work exists, we compare our method against a baseline of always localizing to the whole image (21.99% accuracy), and compare our models using different detectors and localization strategies.

As reference, we also compare our model against supervised approaches trained in a fully [3, 11, 23, 29] or weakly [1, 35] supervised setting. Note that these systems are not directly comparable to ours. In fact, the comparison is unfavourable to us as these work also use external tools like visual detectors or bounding box proposal generators in addition to supervised phrase localization training data.

Table 1 shows the accuracies on Flickr30kEntities for bounding box predictions from a selection of our models, using different combination of detectors, concept selection and localization strategies. Our best performing model combines **tfcoco**, **tfoid** and **places365** detectors with the **w2v-max** concept selector and the **union** localization strategy. This model comfortably outperformed the state-of-the-art weakly supervised model [1] on this dataset (50.49% vs. 38.71%). Its accuracy is also higher than a strongly supervised model [29] (47.81%) and is competitive against others [3, 11, 23] that use strong supervision along with specialised detectors for the dataset, part-of-speech taggers and parsers, the full caption, and takes into account other entities/relations mentioned in the caption. In contrast, our method is much simpler and does not rely on domain-specific paired training data. The full results with different detector combinations, concept selection and localization strategies are provided as supplementary material. These results suggest that paired annotations might not even be completely necessary for the task, at least for Flickr30kEntities.

Detector	**Similarity**	**Strategy**	**Acc (UB) %**
Baseline: Always localize to whole image			21.99
CC+OI	-	largest	30.32 (73.00)
20	w2v-avg	union	36.49 (51.81)
CC	w2v-max	union	37.57 (51.22)
OI	w2v-max	union	44.69 (50.04)
CC+OI	w2v-max	union	48.20 (55.85)
CC+OI+PL+CL	w2v-avg	consensus	49.51 (58.93)
CC+OI+PL+CL	w2v-avg	union	49.61 (58.10)
CC+OI+PL	w2v-avg	consensus	50.11 (58.00)
CC+OI+PL+CL	w2v-last	union	50.36 (57.81)
CC+OI+PL	w2v-max	union	**50.49** (57.81)
Weakly supervised			
GroundeR [29]			28.94
Yeh *et al.* [35]			36.93
KAC Net + Soft KBP [1]			38.71
Strongly supervised			
GroundeR [29]			47.81
SPC+PPC [23]			55.85
QRC Net [3]			65.14
Query Adaptive R-CNN [11]			65.21

Table 1. Accuracies (and upperbound **UB**) of some of our selected models on Flickr30kEntities, comparing different detector combinations, semantic similarity measures and localization strategies. As a comparison against supervised settings, we present our results alongside selected strongly and weakly supervised systems. These systems are not directly comparable to ours as they use phrase localization annotations for training. Keys: CC=**tfcoco**, OI=**tfoid**, 20=**tfcoco20**, PL=**places365**, CL=**colour**.

Table 2 gives the per-category breakdown of the accuracies. Our best models resulted in higher accuracies than all strongly supervised models for two out of eight categories (animals and vehicles). Our models also achieved better accuracies than weakly supervised models in seven out of eight categories, and are competitive for the remaining category (scene) against KAC Net [1] (40.58% vs. 43.53%) and outperformed Yeh *et al.* [35] (24.87%).

[5] Our evaluation script can be found at https://github.com/josiahwang/phraseloceval.

	people	**clothing**	**bodyparts**	**animals**	**vehicles**	**instruments**	**scene**	**other**	**overall**
# instances	5626	2306	523	518	400	162	1619	3374	14481
20 (max, u)	60.31	9.63	2.10	82.43	74.75	19.14	17.85	17.96	36.33
CC (max, u)	56.35	10.45	1.72	83.59	79.25	17.90	15.69	29.79	37.57
CC+OI (max, u)	*66.34*	*37.99*	21.03	**84.75**	79.75	*47.53*	20.14	33.11	48.20
CC+OI+PL (avg, u)	66.18	35.52	21.03	**84.75**	**81.00**	*47.53*	39.16	*34.71*	50.27
CC+OI+PL (max, u)	66.27	37.55	20.65	**84.75**	80.00	*47.53*	38.91	34.41	50.49
CC+OI+PL+CL (avg, u)	65.22	35.65	*21.22*	78.19	78.00	*47.53*	40.58	34.05	49.61
Weakly supervised									
Yeh *et al.* [35]	58.37	14.87	2.29	68.91	55.00	22.22	24.87	20.77	20.91
KAC Net (Soft KBP) [1]	58.42	7.63	2.97	77.80	69.00	20.37	43.53	17.05	38.71
Strongly supervised									
SPC+PPC [23]	71.69	50.95	25.24	76.25	66.50	35.80	51.51	35.98	55.85
QRC Net [3]	76.32	59.58	25.24	80.50	78.25	50.62	67.12	43.60	65.14
Query Adaptive R-CNN [11]	78.17	61.99	35.25	74.41	76.16	56.69	68.07	47.42	65.21

Table 2. Non-paired phrase localization accuracies for different phrase types, as defined in Flickr30kEntities. **Bolded** results show higher accuracies than strongly supervised models, while *italicized* accuracies indicate that they are higher than weakly supervised models. Keys: CC=**tfcoco**, OI=**tfoid**, 20=**tfcoco20**, PL=**places365**, CL=**colour**, max=**w2v-max**, avg=**w2v-avg**, u=**union**.

4.2.1 Discussion

Upperbound. The upperbound accuracies generally increase as we increase the number of detectors and categories used. This indicates that the recall has increased, presumably by virtue of more candidate bounding boxes being proposed. Our concept selection process reduces this upperbound, but as a result allows the localization strategy to perform the task more accurately. Interestingly, the upperbound did not change significantly when only a subset of 20 categories of **tfcoco** is used and with concept selection applied (51.81% vs. 51.22% in Table 1). This is because of the large number of people-related phrases in the dataset; the *person* detectors in both detector groups manage to capture this.

Detector. The accuracy generally improves with more detectors (and number of categories), as long as the detections are of high quality. While the differences between **tfcoco20** (20 categories) and **tfcoco** (80 categories) are much smaller, using **tfoid** (545 categories) resulted in larger improvements (see Figure 3). Detector quality is also important, as demonstrated by the generally weak performance from **yolo9000** (accuracies are generally lower than 20%) which has a low accuracy (19.7 mAP on a subset of 200 categories [27]) despite boasting the ability to detect over $9,000$ categories. The category labels themselves include abstract categories (*thing*, *instrumentation*), which are irrelevant as these are not often used to describe objects. The **colour** detectors on their own gave low accuracies (generally <10%). This is because only a small subset of test phrases contained colour terms, and the connected component labelling also resulted in generally small bounding boxes; however, using the union of bounding boxes for localization resulted in better accuracies (≈18%).

Combination of detectors. The detectors are also complementary to each other. Combining **tfcoco** and **tfoid** results in a higher accuracy (48.20%) than using either alone (37.57% and 44.69% respectively). From Table 2, we observe that **tfoid** helped improve over **tfcoco** especially for *clothing* (by ≈27% accuracy), *bodyparts* (≈20%) and *instruments* (≈30%), and to a certain extent *scenes*. It also provided some additional redundancy to help localize *person* since it contains different people detectors (*person*, *man*, *woman*, *boy*, *girl*). **places365** improved the localization of scene phrases (≈19%).

Colour detector. Adding a **colour** detector to **tfcoco+tfoid+places365** does not improve the overall accuracies (CC+OI+PL (avg, u) vs. CC+OI+PL+CL (avg, u) in Table 2), but it helps with scene-type phrases, especially when the scene contains a colour term and covers most of the image. It also helps when the phrase is a single colour noun (*red*), when the head noun is not detected (*an orange outfit*), or when the colour can be inferred for a missed detection (*tree*). This works as long as there are no other objects with the same colour. Some problematic cases are when the desired colour occurs elsewhere in the image, and with phrases such as *a white man*. Figure 4 shows some examples illustrating the contributions of the **colour** detector. We further quantitatively investigate the contributions of **colour** to **tfcoco+tfoid+places365** by evaluating on a subset of test phrases where the 11 basic colour terms occur (Table 3). We observe that colour terms are most frequently mentioned in *clothing*-type phrases. Adding a **colour** detector improves localization in *clothing* and *scene* phrases.

Concept selection. Our concept selection process with word embeddings similarity is intuitive, and results in accurate localization. **tfcoco** performed on par with Yeh *et*

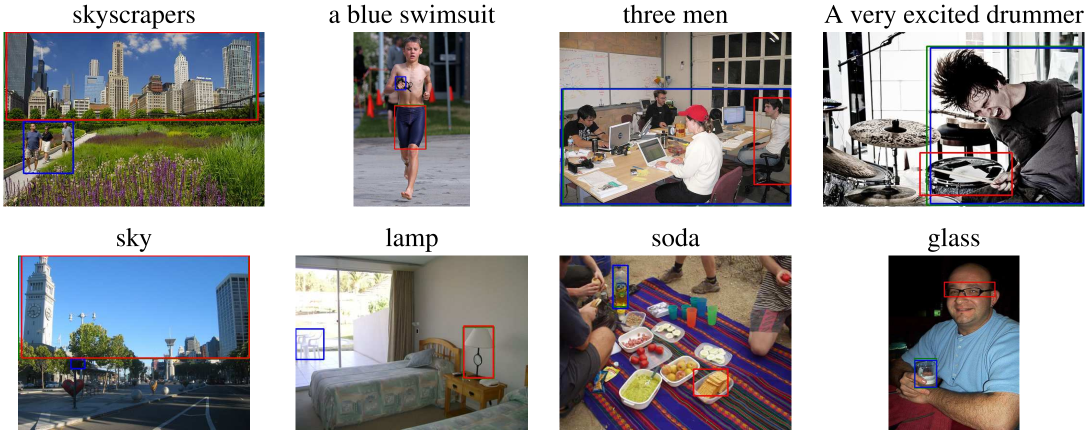

Figure 3. Example localization output for Flickr30kEntities (top row) and ReferItGame (bottom row). We compare the effects of adding a **tfoid** detector (red bounding box) to **tfcoco** (blue) (**w2v-max**, **union**). The ground truth is indicated in green. The first two columns show examples of where adding a **tfoid** detector improves localization, while the last two columns are examples where it hurts localization.

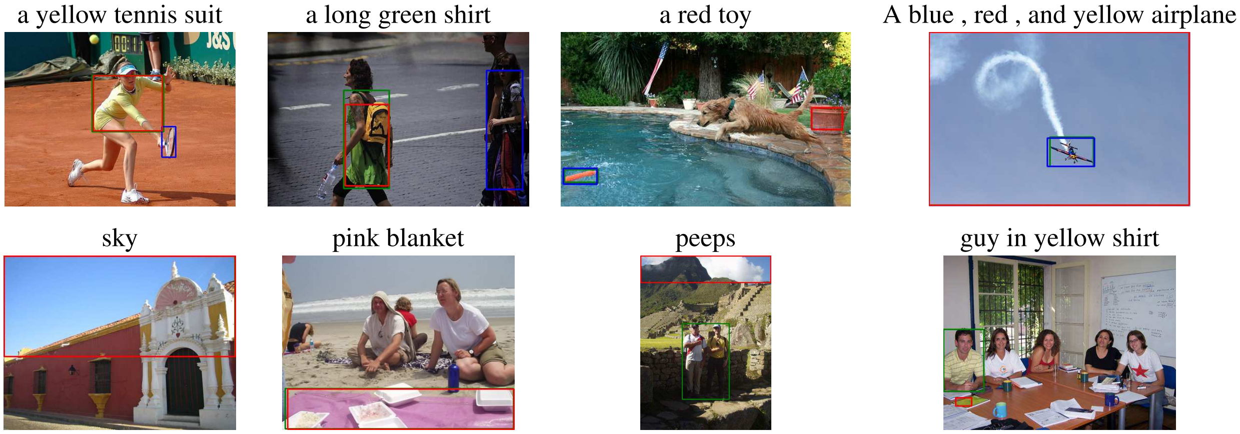

Figure 4. Example localization output for Flickr30kEntities (top row) and ReferItGame (bottom row). We compare the effects of adding a **colour** detector (red bounding box) to **tfcoco+tfoid+places365** (blue) (**w2v-avg**, **union**). The ground truth is indicated in green. The first two columns show examples of where adding a **colour** detector improves localization, while the last two columns are examples where it hurts localization.

al. [35] which computes the similarity using paired annotations, and with the same 80 categories. Our approach captures distributional similarities, and is undesirable in certain cases, for example a *cyclist* is more similar to a *bicycle* or a *wheel* than it is to a *person*. We also found that all three word vector aggregation schemes perform comparably; **w2v-last** generally performs similar to **w2v-max** with only a minor degradation. This agrees with our assumption that the last word in the phrase is most likely the head word in Flickr30kEntities. **w2v-avg** also performs very slightly worse in general, except when **colour** detectors are used.

Localization strategy. For this dataset, the **union** localization strategy seems to work best, partly because of how the test dataset is constructed. It is also useful for **colour** detectors which produce generally small bounding boxes. The **largest** strategy also works reasonably well; objects mentioned in the captions tend to be larger than those not. Our novel **consensus** strategy, designed to allow for slightly higher upperbounds and accuracies by voting, generally gave accuracies comparable to **union**-based equivalents.

4.3. Phrase localization on ReferItGame

ReferItGame [15] crowdsources phrases from an interactive game to describe segments in IAPR TC-12 images [9]. It is significantly different from Flickr30kEntities as phrases are not extracted from image captions and are also much shorter. We use the test split of Rohrbach *et al.* [29] consisting of $65,193$ phrases for $9,999$ images[6]. Again, the training and validation splits are ignored.

Table 4 shows the accuracies on ReferItGame for a selected set of our models, again with different combinations of detectors, concept selection and localization strategies.

[6] We used the split provided at https://github.com/lichengunc/refer

	people	clothing	bodyparts	animals	vehicles	instruments	scene	other	overall
# instances	30	1323	48	177	79	1	93	292	2033
CC+OI+PL (avg, u)	66.67	34.24	29.17	89.27	79.75	100.00	44.09	48.63	43.43
CC+OI+PL+CL (avg, u)	40.00	35.75	27.08	69.49	65.82	100.00	58.06	44.86	41.81

Table 3. Phrase localization accuracies for different phrase types on a subset of query phrases that contain at least one basic colour term.

Detector	Similarity	Strategy	Acc (UB) %
Baseline: Always localize to whole image			14.64
20	w2v-max	largest	14.97 (26.82)
CC	w2v-max	largest	15.40 (27.16)
OI	w2v-max	largest	19.82 (28.03)
CC+OI	w2v-avg	largest	21.21 (32.70)
CC+OI+PL	w2v-avg	consensus	22.25 (35.56)
CC+OI+PL	w2v-avg	largest	23.95 (35.04)
CC+OI+PL+CL	w2v-max	consensus	25.52 (42.48)
CC+OI+PL+CL	w2v-max	largest	**26.48** (39.50)
Weakly supervised			
GroundeR [29]			10.70
KAC Net + Soft KBP [1]			15.83
Yeh *et al*. [35]		yolococo	17.96
Yeh *et al*. [35]		vgg+yolococo	20.91
Strongly supervised			
GroundeR [29]			26.93
Hu *et al*. [12]			27.80
QRC Net [3]			44.07

Table 4. Accuracies of some of our selected models on ReferItGame. Again, we present our results alongside selected strongly and weakly supervised systems as a comparison since no previous non-paired model exists. Keys: CC=**tfcoco**, OI=**tfoid**, 20=**tfcoco20**, PL=**places365**, CL=**colour**.

The accuracy of the baseline of always localizing to the whole image is 14.64%. Our best performing model again performs better than all weakly supervised models (26.48% vs. the state of the art's 20.91%), and is on par with some strongly supervised models [12, 29], although not at the level of QRC Net [3].

4.3.1 Discussion

Localization strategy. Unlike Flickr30kEntities, taking the **union** does not perform as well as simply taking the **largest** box; this is consistent across models. Again, our proposed **consensus** strategy performs well, although not generally as well as the **largest** strategy.

Concept selection. Like Flickr30kEntities, **w2v-max** and **w2v-avg** performs equally well, with **w2v-max** having a very slight edge. Unlike Flickr30kEntities, **w2v-last** performs substantially worse than other semantic similarity measures. This is because the phrases are short, and the head word is more likely to be mentioned at the beginning.

Detectors. The detectors generally show a similar behaviour as with Flickr30kEntities. Adding **tfoid** to **tfcoco** pushed the accuracy beyond the state of the art [35], and adding **places365** further increased its accuracy. Unlike Flickr30kEntities, the **colour** detector contributed substantially more, increasing the overall accuracy by ≈3%. ReferItGame has many colour-based phrases (many single colour words), due to how the annotations were obtained. The model also performs well by inferring the colour of *sky*, *cloud* and *tree*, which occur frequently (Figure 4).

4.4. Phrase localization on Visual Genome

To demonstrate our model's scalability to different datasets, we also test our model on Visual Genome [17] where the queries are at sentence level, rather than at phrase level. Zhang *et al*. [39] reported 26.4% localization accuracy with a fully supervised method. The only weakly supervised equivalent of which we are aware reported 24.4% accuracy [34], but this is an unfair comparison because they evaluated whether a single point falls inside the bounding box, rather than predicting the full box. Our model (**tfoid**, **w2v-max**, **largest**) achieved 14.29% accuracy on Visual Genome, and by combining **tfoid** with **tfcoco** the accuracy is increased to 16.39%. This observation is consistent to what we reported, and we infer that the same pattern should apply to further combinations and variants.

5. Conclusions

We introduced the first approach to phrase localization in images without phrase localization annotations. This non-paired approach, while simple, proved effective: In experiments with Flickr30kEntities and ReferItGame it outperformed all existing weakly supervised approaches and performed competitively to strongly supervised approaches. The method is a strong baseline – phrase localization can be successfully performed on these datasets even without paired examples. Our work suggests that there is significant room for simpler and general methods that rely on few/no paired annotations, instead of complex models that attempt to fit paired annotations to achieve high performance improvements without the ability to generalize. This finding can change how Language & Vision tasks are viewed and tackled in future – researchers should make better use of paired annotations beyond what can already be achieved without such task-specific data.

References

[1] Kan Chen, Jiyang Gao, and Ram Nevatia. Knowledge aided consistency for weakly supervised phrase grounding. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 4042–4050, Salt Lake City, UT, USA, June 2018. IEEE. 1, 2, 4, 5, 6, 8

[2] Kan Chen, Rama Kovvuri, Jiyang Gao, and Ram Nevatia. MSRC: Multimodal spatial regression with semantic context for phrase grounding. *International Journal of Multimedia Information Retrieval*, 7(1):17–28, Mar. 2018. 1, 2

[3] Kan Chen, Rama Kovvuri, and Ram Nevatia. Query-guided regression network with context policy for phrase grounding. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 824–832, Venice, Italy, Oct. 2017. IEEE. 1, 2, 5, 6, 8

[4] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015. 2

[5] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. 3

[6] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998. 4

[7] Ross Girshick. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, Santiago, Chile, Dec. 2015. IEEE. 1, 2

[8] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 580–587, Columbus, OH, USA, June 2014. IEEE. 1, 2

[9] Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. The IAPR TC-12 benchmark: A new evaluation resource for visual information systems. In *International Workshop on Language Resources for Content-Based Image Retrieval, OntoImage'2006*, pages 13–23, Genoa, Italy, May 2006. 7

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1904–1916, Sept. 2015. 1, 2

[11] Ryota Hinami and Shin'ichi Satoh. 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*, pages 2605–2615, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. 1, 2, 5, 6

[12] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural Language Object Retrieval. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 4555–4564, Las Vegas, NV, USA, June 2016. IEEE. 1, 2, 8

[13] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, and Kevin Murphy. Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 3296–3297, Honolulu, HI, USA, July 2017. IEEE. 3

[14] Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1889–1897. Curran Associates, Inc., 2014. 2

[15] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. 1, 2, 7

[16] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. OpenImages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2017. 3

[17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, May 2017. 1, 2, 8

[18] 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 David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755. Springer International Publishing, Sept. 2014. 3

[19] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single shot multibox detector. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 21–37, Amsterdam, The Netherlands, Oct. 2016. Springer International Publishing. 1, 2

[20] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 11–20, Las Vegas, NV, USA, June 2016. 1, 2

[21] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger,

editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013. 4

[22] Bryan A. Plummer, Paige Kordas, M. Hadi Kiapour, Shuai Zheng, Robinson Piramuthu, and Svetlana Lazebnik. Conditional image-text embedding networks. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 258–274, Munich, Germany, Sept. 2018. Springer International Publishing. 1, 2

[23] 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 (ICCV)*, pages 1946–1955, Venice, Italy, Oct. 2017. IEEE. 1, 2, 5, 6

[24] Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k Entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2641–2649, Santiago, Chile, Dec. 2015. 1, 2, 5

[25] 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. *International Journal of Computer Vision*, 123(1):74–93, May 2017. 2, 5

[26] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 779–788, Las Vegas, NV, USA, June 2016. IEEE. 1, 2

[27] Joseph Redmon and Ali Farhadi. YOLO9000: Better, faster, stronger. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 6517–6525, Honolulu, HI, USA, July 2017. IEEE. 1, 2, 3, 6

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc., 2015. 1, 2, 3

[29] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 817–834, Amsterdam, The Netherlands, 2016. Springer International Publishing. 1, 2, 4, 5, 7, 8

[30] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, Dec. 2015. 2, 3

[31] Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, and Yann LeCun. OverFeat: Integrated recognition, localization and detection using convolutional networks. In *Proceedings of the International Conference on Learning Representation (ICLR)*, Banff, Canada, Apr. 2014. 1, 2

[32] Joost van de Weijer, Cordelia Schmid, and Jakob Verbeek. Learning color names from real-world images. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 1–8, Minneapolis, MN, USA, June 2007. IEEE. 3

[33] Mingzhe Wang, Mahmoud Azab, Noriyuki Kojima, Rada Mihalcea, and Jia Deng. Structured matching for phrase localization. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 696–711, Amsterdam, The Netherlands, 2016. Springer International Publishing. 1, 2

[34] Fanyi Xiao, Leonid Sigal, and Yong Jae Lee. Weakly-supervised visual grounding of phrases with linguistic structures. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 5253–5262, Honolulu, HI, USA, July 2017. 1, 2, 8

[35] Raymond A. Yeh, Minh N. Do, and Alexander G. Schwing. Unsupervised textual grounding: Linking words to image concepts. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 6125–6134, Salt Lake City, UT, USA, June 2018. 1, 2, 4, 5, 6, 7, 8

[36] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, Feb. 2014. 5

[37] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling Context in Referring Expressions. In *14th European Conference in Computer Vision (ECCV 2016)*, pages 69–85, Amsterdam, The Netherlands, Oct. 2016. 1, 2

[38] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press, Sept. 2016. 3

[39] Yuting Zhang, Luyao Yuan, Yijie Guo, Zhiyuan He, I-An Huang, and Honglak Lee. Discriminative bimodal networks for visual localization and detection with natural language queries. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 1090–1099, Honolulu, HI, USA, July 2017. 1, 2, 8

[40] 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 & Pattern Recognition (CVPR)*, pages 5696–5705, Salt Lake City, UT, USA, June 2018. 1, 2

[41] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, June 2018. 3