MSRC: Multimodal Spatial Regression with Semantic Context for Phrase Grounding

Conference Paper · June 2017

DOI: 10.1145/3078971.3078976

Kan Chen
University of Southern California

MSRC: Multimodal Spatial Regression with Semantic Context for Phrase Grounding

Kan Chen
University of Southern California
kanchen@usc.edu

Rama Kovvuri
University of Southern California
nkovvuri@usc.edu

Jiyang Gao
University of Southern California
jiyangga@usc.edu

Ram Nevatia
University of Southern California
nevatia@usc.edu

ABSTRACT

Given an image and a natural language query phrase, a grounding system localizes the mentioned objects in the image according to the query's specifications. State-of-the-art methods address the problem by ranking a set of proposal bounding boxes according to the query's semantics, which makes them dependent on the performance of proposal generation systems. Besides, query phrases in one sentence may be semantically related in one sentence and can provide useful cues to ground objects. We propose a novel Multimodal Spatial Regression with semantic Context (MSRC) system which not only predicts the location of ground truth based on proposal bounding boxes, but also refines prediction results by penalizing similarities of different queries coming from same sentences. The advantages of MSRC are twofold: first, it removes the limitation of performance from proposal generation algorithms by using a spatial regression network. Second, MSRC not only encodes the semantics of a query phrase, but also deals with its relation with other queries in the same sentence (*i.e.*, context) by a context refinement network. Experiments show MSRC system provides a significant improvement in accuracy on two popular datasets: Flickr30K Entities and Refer-it Game, with 6.64% and 5.28% increase over the state-of-the-arts respectively.

CCS CONCEPTS

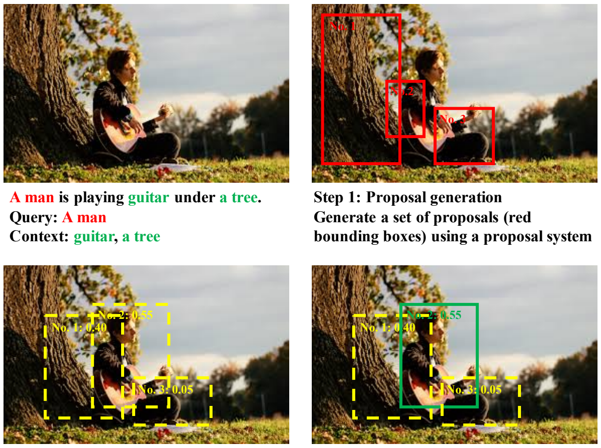
•**Computing methodologies** →**Visual content-based indexing and retrieval;** *Information extraction;* Scene understanding;

KEYWORDS

phrase grounding; spatial regression; multimodal; context

ACM Reference format:
Kan Chen, Rama Kovvuri, Jiyang Gao, and Ram Nevatia. 2017. MSRC: Multimodal Spatial Regression with Semantic Context for Phrase Grounding. In *Proceedings of ICMR '17, June 6–9, 2017, Bucharest, Romania, ,* 9 pages.
DOI: http://dx.doi.org/10.1145/3078971.3078976

1 INTRODUCTION



A man is playing guitar under a tree.
Query: A man
Context: guitar, a tree

Step 1: Proposal generation
Generate a set of proposals (red bounding boxes) using a proposal system

Step 2: Multimodal Spatial Regression
Language input: "A man"
Regress proposals based on query's semantic and visual features

Step 3: Context refinement
Language input: "A man", "guitar", "a tree"
Refine choice of regression boxes by parsing context information

Figure 1: Multimodal Spatial Regression with semantic Context (MSRC) system regresses each proposal based on query's semantics and visual features. Besides, MSRC takes advantage of context cues to filter out confusing candidates and refine regression results. (Each regression box's ID corresponds to proposal box's ID, with confidence on the top-left corner.)

Given an image and a natural language query phrase, phrase grounding attempts to localize the mentioned objects in the image according to the query's specification. It can be utilized in many daily life applications, such as electronic entertainment, early education, security surveillance, *etc.* Solution to this problem can be an important building block for image-language related tasks, such as image captioning [1, 6, 14], visual question answering [2, 3, 7] and image retrieval [10, 24].

Phrase grounding is a challenging problem in reasoning language queries and transferring their semantics to localize object in visual contents. To address this problem, typically a set of proposal bounding boxes are first generated as candidates by some proposal generation system. The main difficulties then lie in how to correlate language input and proposals' features and how to localize

ICMR '17, June 6–9, 2017, Bucharest, Romania
© 2017 ACM. ACM ISBN 978-1-4503-4701-3/17/06...$15.00
DOI: http://dx.doi.org/10.1145/3078971.3078976

objects after learning such multimodal correlation. State-of-the-art methods address the first difficulty by treating phrase grounding as a ranking problem, and learn a multimodal subspace where relevance between visual and language inputs are measurable and then rank proposals according to the relevance of query's specification. Among these, Phrase-Region CCA [23] and SCRC [13] models learn a subspace using Canonical Correlation Analysis (CCA) and a Recurrent Neural Network (RNN) respectively. GroundeR [27] adopts an attention network, which learns a latent subspace to attend on related proposals given different queries via phrase reconstruction.

These methods are bounded by two limitations. First, when the proposal generation system fails to provide good proposals which overlap the object mentioned by the query with a large region (Fig. 1), these proposal-based models are unable to localize the correct object; as a result, there exists a performance upper bound from proposal generation systems. Second, these grounding systems consider query phrases as unrelated to each other; however, sometimes different queries for the same image are semantically related. For example, we observe query phrases for the same images are usually correlated in same image-related captions in Flickr30K Entities [23] dataset. Intuitively, given a query phrase, other phrases from the same sentence, which are the query's context, can provide useful cues for grounding the correct objects. As shown in Fig. 1, with the context "the tree", the grounding system should be able to infer the current query "a man" does not refer to the tree object in the image, even though the proposal of tree also has high confidence; and the context "guitar" can provide hints for the system to find "a man" near the object "guitar".

To address the aforementioned two issues, we propose a Multimodal Spatial Regression with semantic Context (MSRC) deep network. MSRC system is composed of two parts: a Spatial Regression Network (SRN) and a Context Refinement Network (CRN). SRN applies a bi-directional LSTM to encode the query and takes each proposal's feature as visual information. It learns to predict the probability of each proposal being related to query, and regresses each proposal to the mentioned object's location via a joint projection in a multimodal subspace. SRN is robust to performance of proposal generation program, because when proposal does not overlap much with the mentioned object in the query, SRN can regress the proposal to best fit the mentioned object. CRN takes a pair of queries from the same sentence and jointly predicts proposals' probability of being relevant to each query. Based on the assumption that different queries in the same sentence refer to different objects, CRN adopts a joint prediction loss, which penalizes the probabilities of confusing proposal candidates. The final prediction of MSRC is by a late fusion of SRN decision and CRN decision, which takes advantages of both spatial regression results and context information.

We evaluate the MSRC system on two popular phrase grounding datasets: Flickr30K Entities [23] and Refer-it Game [15] datasets. Flickr30K Entities contains more than 30K images associated with 5 captions for each image. There are 244K query phrases referring to 276K manually annotated bounding boxes of objects in images. Every query phrase comes from some image related caption. Referit Game has 130K query phrases, referring to 96K objects, which are annotated for 19K images of natural scenes. We adopt ratio of phrases which are successfully grounded by MSRC as metric. Experiments show that MSRC system has more than 6% improvement in Flickr30K Entities and 5% improvement in Refer-it Game datasets, indicating the effectiveness of our approach.

Our contributions are two fold: First, we propose a spatial regression approach in multimodal space, which relieves performance limitation from proposal generation systems. Second, we encode context information with query phrase by adopting a joint prediction loss during training stage, which helps filter out confusing candidates during grounding. In the following paper, we first discuss related work in phrase grounding problem in Sec. 2. More details about MSRC system are presented in Sec. 3. Finally, experiments of MSRC system and related results are provided and analyzed in Sec. 4.

2 RELATED WORK

Language grounding. Language grounding is a hot topic in both computer vision and natural language processing communities. To parse the semantic information from queries, Krishnamurthy *et al.* [19] and Matuszek *et al.* [21] introduce joint learning models to parse knowledge in language sentences and apply models on scene understanding, attribute classification and geographical question answering. Recently, Wang *et al.* [30] propose a structured system to match phrase and deal with the "partial match coreference" relation in queries to boost performance. To correlate visual contents with language embeddings, Karpathy *et al.* [17] propose to align sentence fragments and image regions in a subspace, and replace the dependency tree with a bi-directional RNN in [1]. Plummer *et al.* [23] apply a Canonical Correlation Analysis (CCA) model to ground object in images. Based on these methods, Hu *et al.* [13] propose a hierarchy of RNNs to retrieve objects given a query. Rohrbach *et al.* [27] learn to attend on mentioned object via phrase reconstruction in the unsupervised scenario. These methods learn the correlation between language and visual modalities, and are effective when a proposal system generates good candidate proposals. However, their performances are dependent on the proposal system's upper bound. As a result, these methods are unable to localize the mentioned object when there are no proposals overlapping much with it.

Spatial regression. Spatial regression is successfully applied in object detection. Fast R-CNN [8] first introduces regression loss in object detection. Based on this, Ren *et al.* [26] adopt a Region Proposal Network (RPN) which further improves the accuracy in proposal regression. Redmon *et al.* [25] regress image in grid level and merge regressed grids using the non-max suppression algorithm, which efficiently improves the detection speed. Liu *et al.* [20] integrate proposal generation in a single network and discretize the output into a set of bounding boxes with default sizes over different ratios and scales of feature maps, which further improve the accuracy and speed of detection. Inspired by the success of regression in object detection, MSRC system applies a spatial regression network in phrase grounding problem, which is robust to different proposal generation algorithms and improves the performance of grounding.

Semantic context encoding. Context brings useful information and is successfully applied in image-sentence referring and object detection problems. Kantorov *et al.* [16] propose a ContextLocNet which encodes region proposals and context around regions

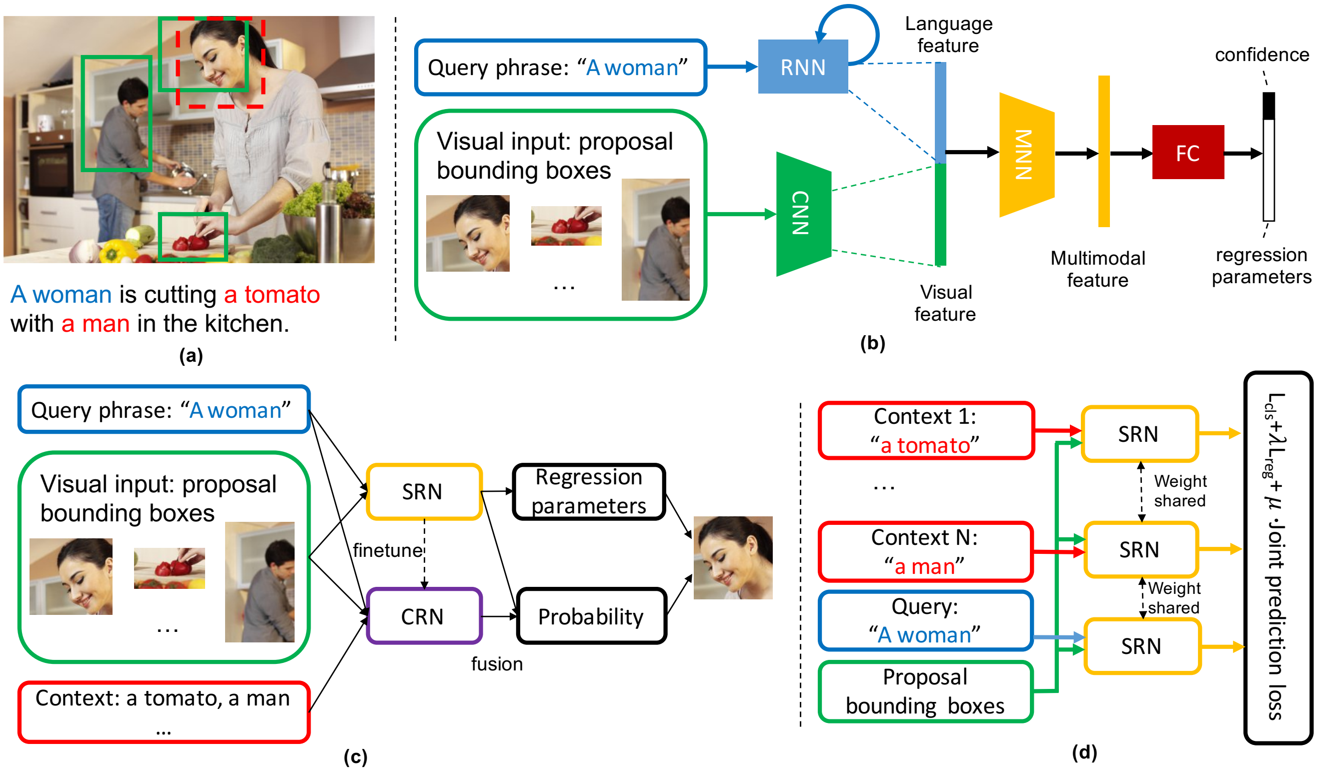


Figure 2: Structure of MSRC system. (a) An example image and query phrases: For query "A woman" (blue text), queries in red text are considered as its context, which are further utilized by CRN. Input image is represented as a set of proposal bounding boxes (green), and the ground truth for the query is the red box. (b) Structure of SRN: SRN takes proposals and query phrase as inputs. Multimodal features are encoded by a Multimodal Neural Network (MNN). SRN predicts each proposal's probability of being related to the query as well as regression parameters to localize the mentioned object. (c) Framework of MSRC: A SRN is first trained and utilized to finetune CRN later. CRN refines probability predicted by SRN via encoding context information. (d): Structure of CRN: Each (language, proposal set) pair has a SRN to predict confidence. All SRNs share weights during training. We propose a joint prediction loss to encode context information.

using context-aware models, and achieve the state-of-the-art performance on PASCAL VOC 2007 dataset [5]. Nagaraja *et al.* [22] introduce a Multiple Instance Learning (MIL) approach to the object referring problem, which learns to rank query regions among context regions in images to boost performance. Yu *et al.* [31] jointly predict all query regions in one image, and encode context information by sharing information between CNNs and RNNs, which achieves the state-of-the-art performance in referring task. MSRC system is inspired by the success of using context information. However, it considers language phrases rather than visual contents as context information, and applies a context refinement network on grounding problem.

3 MSRC SYSTEM

Multimodal Spatial Regression with semantic Context (MSRC) system contains two parts: a Spatial Regression Network (SRN) and a Context Refinement Network (CRN). We first introduce the framework of MSRC followed by structures of SRN and CRN. Then we provide more details about training and grounding of MSRC system.

3.1 Framework

The visual input of SRN and CRN is a set of N proposal bounding boxes $\{r_i\}$ generated from an input image I. Each proposal r_i is represented as the visual feature $\mathbf{x}_i \in \mathbb{R}^{d_v}$ extracted by a pre-trained Convolutional Neural Network (CNN), where d_v is the visual feature's dimension. The language input of SRN is a query phrase q, while CRN takes q and its context phrases $\{p_i^q\}$ parsed from the same sentence as inputs. We adopt a Bi-directional LSTM [12] to encode semantics of query q and its context $\{p_i^q\}$, which are denoted as $\mathbf{q} \in \mathbb{R}^{d_l}$ and $\mathbf{p}_i^q \in \mathbb{R}^{d_l}$ respectively, where d_l is the language embedding vector's dimension. In this paper, we select query phrases from the same image description of query $\mathbf{q}$ as its context phrases $\{\mathbf{p}_i^q\}$.

Given $\{\mathbf{x}_i\}$ and $\mathbf{q}$, SRN predicts a probability distribution over the N proposals as well as each proposal's regression parameters to infer the location of object mentioned by the query phrase q. The objective is defined as:

$$\arg\min_{\theta_s} \sum_{\mathbf{q}} \left[L_{cls}^s(\{x_i\},\mathbf{q}) + \lambda L_{reg}^s(\{\mathbf{x}_i\},\mathbf{q})\right] \quad (1)$$

where θ_s is the parameters of SRN. L^s_{cls} is a cross entropy function adopted for multiclass classification task. L^s_{reg} is a regression loss function, weighted by a hyper parameter λ, whose details are in Sec. 3.2

To further refine the grounding results, we propose a CRN to generate a more accurate probability distribution over the N proposals via encoding context information. CRN is based on an assumption: different phrases in one sentence refer to different objects in one image, which is always true in Flickr30K Entities dataset [23]. Given N proposals' visual features $\{\mathbf{x}_i\}$, language features of query q and M context phrases $\{p^q_j\}$, the objective of CRN is defined as:

$$\arg\min_{\theta_c} \sum_{\mathbf{q}} \left[L^c_{cls} + \lambda L^c_{reg} + \mu J(\{\mathbf{x}_i\}, \mathbf{q}, \{\mathbf{p}^q_j\}) \right] \tag{2}$$

where θ_c is the parameters of CRN and μ, λ are hyper parameters. L^c_{cls} and L^c_{reg} are similar to L^s_{cls} and L^s_{reg} in Eq. (1). We propose a novel joint prediction loss function J, penalizing prediction results which match the semantic of context phrases $\{p^q_i\}$. With this term, CRN penalizes the probabilities of proposals referred by context queries in the same sentence.

We first train SRN and then finetune CRN based on the trained weights of SRN. During evaluation, we fuse the predicted probabilities from both SRN and CRN, and regress the proposal with maximum probability according to the regression parameters predicted by SRN, with more details in Sec. 3.4.

3.2 Spatial Regression Network (SRN)

As shown in Fig. 2, SRN concatenates language embedding vector $\mathbf{q}$ with each of the proposal's visual feature $\mathbf{x}_i$. It then applies a network to generate multimodal features $\{\mathbf{v}^q_i\} \in \mathbb{R}^m$ for each of the $\langle q, r_i \rangle$ pair in a m-dimensional subspace (We term the network as MNN). The multimodal feature $\mathbf{v}^q_i$ is calculated as:

$$\mathbf{v}^q_i = \varphi(\mathbf{W}_m(\mathbf{q}||\mathbf{x}_i) + \mathbf{b}_m) \tag{3}$$

where $\mathbf{W}_m \in \mathbb{R}^{(d_l+d_v)\times m}, \mathbf{b}_m \in \mathbb{R}^m$ are projection parameters of MNN. $\varphi(.)$ is a non-linear activation function. "||" denotes a concatenation operator. We also replace MNN with a Multimodal Compact Bilinear Pooling layer in [7] to evaluate performances of different multimodal features, with more details discussed in Sec. 4.

Given the multimodal feature $\mathbf{v}^q_i$, SRN predicts a 5D vector $\mathbf{s}^p_i \in \mathbb{R}^5$ for each proposal r_i via a linear projection (superscript "p" denotes prediction).

$$\mathbf{s}^p_i = \mathbf{W}_s \mathbf{v}^q_i + \mathbf{b}_s \tag{4}$$

where $\mathbf{W}_s \in \mathbb{R}^{d\times 5}$ and $\mathbf{b}_s \in \mathbb{R}^5$ are projection weight and bias to be optimized. The first element in $\mathbf{s}^p_i$ indicates the confidence of r_i being related to input query q's semantics. We denote $\{\mathbf{s}^{p\prime}_i\}$ as the probability distribution of $\{r_i\}$ after we feed $\{\mathbf{s}^p_i[0]\}$ to a softmax function. During training, we choose the positive label of proposal as the one which overlaps most with ground truth and with Intersection of Union (IoU)> 0.5. Thus, the classification loss is calculated as:

$$L^s_{cls}(\{\mathbf{x}_i\}, \mathbf{q}) = -\log(\mathbf{s}^{p\prime}_{i^*}[0]) \tag{5}$$

where i^* is positive proposal's index in the proposal set.

The next four elements of $\mathbf{s}^p_i$ record the regression information based on current location of r_i, which are defined as:

$$\begin{aligned} \mathbf{s}^p_i[1] &= (x_{pred} - x_{r_i})/w_{r_i} \\ \mathbf{s}^p_i[2] &= (y_{pred} - y_{r_i})/h_{r_i} \\ \mathbf{s}^p_i[3] &= \log(w_{pred}/w_{r_i}) \\ \mathbf{s}^p_i[4] &= \log(h_{pred}/h_{r_i}) \end{aligned} \tag{6}$$

where $[x_{pred}, y_{pred}, w_{pred}, h_{pred}]$ are the predicted regressed bounding box's center x, y coordinates, width and height. Similarly, $[x_{r_i}, y_{r_i}, w_{r_i}, h_{r_i}]$ is the location information of r_i.

Each proposal's ground truth regression data $\mathbf{s}^q_i \in \mathbb{R}^4$ is calculated in the same way as Eq.(6), by replacing $[x_{pred}, y_{pred}, w_{pred}, h_{pred}]$ with the ground truth bounding box's location information. The regression loss for SRN is:

$$L^s_{reg}(\{\mathbf{x}_i\}, \mathbf{q}) = \frac{1}{4N} \sum_{i=1}^{N} \sum_{j=0}^{3} f\left(\left| \mathbf{s}^p_i[j+1] - \mathbf{s}^q_i[j] \right| \right) \tag{7}$$

where $f(.)$ is the smooth L1 loss function: $f(x) = 0.5x^2(x < 1)$, and $f(x) = |x| - 0.5(x \geq 1)$.

3.3 Context Refinement Network (CRN)

CRN is built on an assumption: different query phrases in one sentence usually refer to different objects in the same image, which is common in Flickr30K Entities dataset [23]. Based on this assumption, CRN penalizes prediction results which match context's semantics during training. In this way, CRN maximizes probabilities of proposals referred by queries rather than context from same sentences.

CRN takes proposals' visual features $\{\mathbf{x}_i\}$, embedding vector $\mathbf{q}$ of query q, and M context phrases' embedding vectors $\{\mathbf{p}^q_j\}$ as inputs. As shown in Fig. 2, for each query or its context phrase, CRN builds an SRN-like structure to process language embedding and visual features, which share weights with each other during training. We denote the output of SRN for pair $\langle \mathbf{q}, \{\mathbf{x}_i\} \rangle$ as $\{\mathbf{t}^q_i\}$ and for pair $\langle \mathbf{p}_j, \{\mathbf{x}_i\} \rangle$ as $\{\mathbf{t}^{p_j}_i\}$. Besides, we denote the proposal set S_q as the proposals which overlap with ground truth with IoU> 0.5. The cross-entropy loss and regression loss in Eq. (2) are calculated in the same way as Eq. (5) and Eq. (7), while the last term in Eq. (2) is calculated as:

$$J(\{\mathbf{x}_i\}, \mathbf{q}, \{\mathbf{p}^q_j\}) = -\sum_{j=1}^{M} \sum_{i=1}^{N} \delta(i \in S_q) \log(\mathbf{t}^{p_j\prime}_i[0]) \tag{8}$$

where $\mathbf{t}^{p_j\prime}_i[0]$ is the softmax normalized probability generated from $\{\mathbf{t}^{p_j}_i[0]\}$. $\delta(.)$ is an indicator function to judge whether current label matches query q's semantic. If yes, then CRN penalizes these results, because the proposal should belong to context phrase rather than input query.

3.4 Training & Phrase grounding of MSRC

The number of context training samples is much smaller than that of single query phrases. To compensate for this, we first train the SRN using single queries with image proposals, and then we finetune the CRN by introducing context training samples and initialize CRN structure by the pre-trained SRN. We use the Adam algorithm [18] to optimize the deep learning framework, and adopt the rectified linear unit (ReLU) as the non-linear activation function.

During phrase grounding stage, if an input query q has T_q context phrases in the same sentence, SRN first predicts a probability distribution $\{\mathrm{Pr}_{srn}(i)\}$ over the proposals as well as each proposal's regression parameters $\{\mathrm{Reg}_{srn}(i)\}$. Then CRN receives a triplet $\langle q, \{r_i\}, p_j^q \rangle$ successively and predicts a probability distribution over the proposals $\{\mathrm{Pr}_{crn}(i|p_j^q)\}$. MSRC system gives the final prediction result as:

$$\mathrm{Box}^* = \mathrm{Regression}(\mathrm{Reg}_{srn}(i^*), r_i), \text{ where}$$
$$i^* = \arg\max_i \left(\mathrm{Pr}_{srn}(i) + \frac{\tau}{T_q} \sum_{j=1}^{T_q} \mathrm{Pr}_{crn}(i|p_j^q) \right) \tag{9}$$

τ is a hyper parameter. Regression(.) denotes the regression operation based on the input bounding box and regression parameters.

4 EXPERIMENTS

We evaluate MSRC system on Flickr30K Entities [23] and Refer-it Game datasets [15] for phrase grounding.

4.1 Datasets

Flickr30K Entities [23] contains 31,783 images, with 29783, 1000, 1000 images for training, validation and testing respectively. Each image is associated with 5 captions. There are 559,767 query phrases extracted from these captions referring to 276K manually annotated bounding boxes in images. The vocabulary size of queries is 17,150. The maximum length of query phrases is 19 words.

Refer-it Game [15] contains 19,894 images of natural scenes. There are 96,654 distinct objects in these images. Each of them is referred to by 1-3 query phrases (130,525 in total). The vocabulary size of queries is 8800, and the maximum length of query phrases is 19 words.

4.2 Experiment Setup

Proposal generation. We choose Selective Search [29] to generate proposals for Flickr30K Entities and Edge Box [32] to generate proposals for Refer-it Game, which are the same settings as in GroundeR [27] for fair comparison.

Visual feature extraction. We choose a VGG network [28] pretrained on ImageNet [4] to extract each proposal bounding box's visual feature, which is denoted as "$\mathrm{VGG_{cls}}$", for both Flickr30K Entities and Refer-it Game datasets. Besides, we apply a VGG network finetuned by Fast-RCNN [8] on PASCAL VOC 2007 [5] dataset to extract visual features for Flickr30K Entities, which are denoted as "$\mathrm{VGG_{det}}$".

To predict regression parameters, we need to include spatial information from each proposal. For Flickr30K Entities, we augment each proposal's visual feature with its spatial information $[x_{tl}/W, y_{tl}/H, x_{br}/W, y_{br}/H, wh/WH]$ defined in [31]. We denote these augmented features as "$\mathrm{VGG_{cls}}$-SPAT1" and "$\mathrm{VGG_{det}}$-SPAT1" for "$\mathrm{VGG_{cls}}$" and "$\mathrm{VGG_{det}}$" respectively, each being a 4101D ($d_v = 4101$) vector. For Refer-it Game dataset, we augment each proposal's visual feature with its spatial information $[x_{min}, y_{min}, x_{max}, y_{max}, x_{center}, y_{center}, w_{box}, h_{box}]$ defined in [13]. We denote the augmented visual features as "$\mathrm{VGG_{cls}}$-SPAT2", which are 4104D ($d_v = 4104$) vectors.

Approach	Accuracy (%)
Compared approaches	
SCRC [13]	27.80
Wang *et al.* [30]	42.08
GroundeR ($\mathrm{VGG_{cls}}$) [27]	41.56
GroundeR ($\mathrm{VGG_{det}}$) [27]	47.70
MCB [7]	48.69
CCA embedding [23]	50.89
Spatial regression models	
MCB+Reg ($\mathrm{VGG_{det}}$-SPAT1)	51.01
MNN+Reg ($\mathrm{VGG_{cls}}$-SPAT1)	51.18
MNN+Reg ($\mathrm{VGG_{det}}$-SPAT1)	55.99
Context models	
CRN (MNN+Reg($\mathrm{VGG_{det}}$-SPAT1))	56.31
MSRC Full	**57.53**

Table 1: Different models' performance on Flickr30K Entities. CRN is finetuned based on MNN with Regression layer and take $\mathrm{VGG_{det}}$-SPAT1 as input visual features

Language encoding. We encode query's information by a Bi-directional LSTM [12]. We choose the last hidden state from the LSTM as output **q** (dimension $d_l = 1000$), which is same as in [27].

Model initialization. We initialize the convolution layers with the MSRA method [11] and initialize the fully-connected layers with the Xavier method [9]. We introduce batch normalization layers after projecting visual features and language features, which is the same setting as in [27]. During training, we set batch size to be 40, and learning rate to be 0.0005.

Metric. We adopt the accuracy as an evaluation metric, defined to be the ratio of phrases for which the regressed box overlaps with the ground-truth box by more than 50% IoU.

Compared approaches. We choose GroundeR [27], CCA embedding [23], MCB [7] and the approach proposed by Wang *et al.* [30] for comparison, which all achieve state-of-the-art performances in image grounding problem. For GroundeR [27], we focus on its supervised learning scenario, which achieves the best performance among different scenarios.

4.3 Performance on Flickr30K Entities

SRN model. During training, we set the Multimodal Neural Network (MNN) output dimension as 128 ($m = 128$), which is same as [27]. By using the $\mathrm{VGG_{cls}}$-SPAT1 features, we achieve 51.18% accuracy. Compared to GroundeR ($\mathrm{VGG_{cls}}$), SRN achieves 9.62% increase in accuracy. Compared to $\mathrm{VGG_{cls}}$, $\mathrm{VGG_{det}}$ focuses on object detection task, which is more suitable for object localization. By using the $\mathrm{VGG_{det}}$ feature for each proposal, we further improve the performance to 55.99%. We also substitute the MNN with a Multimodal Compact Bilinear pooling (MCB) layer in SRN, which is the model MCB+Reg in Table 1. Experiment shows MCB+Reg has 2.32% increase compared to MCB [7] model.

Phrase Type	people	clothing	body parts	animals	vehicles	instruments	scene	other
GroundeR (VGG_{cls}) [27]	53.80	34.04	7.27	49.23	58.75	22.84	52.07	24.13
GroundeR (VGG_{det}) [27]	61.00	38.12	10.33	62.55	68.75	36.42	58.18	29.08
Wang *et al.* [30]	57.89	34.61	15.87	55.98	52.25	23.46	34.22	26.23
CCA embedding [23]	64.73	46.88	17.21	65.83	68.75	**37.65**	51.39	31.77
SRN: MCB+Reg (VGG_{det}-SPAT1)	62.75	43.67	14.91	65.44	65.25	24.74	64.10	34.62
SRN: MNN+Reg (VGG_{det}-SPAT1)	67.38	47.57	20.11	73.75	72.44	29.34	63.68	37.88
CRN: MNN+Reg (VGG_{det}-SPAT1)	68.24	47.98	20.11	73.94	73.66	29.34	66.00	38.32
MSRC Full	**69.57**	**48.01**	**20.11**	**73.97**	**75.32**	29.34	**66.17**	**39.01**

Table 2: Phrase grounding performances in different phrase types defined in Flickr30K Entities. Accuracy is in percentage.

Multimodal dimension m	64	128	256	512	1024
MNN+Reg	51.21	**55.99**	54.59	55.31	55.97

Table 3: SRN MNN+Reg (VGG_{det}-SPAT1) model's performance (accuracy in %) under different dimension of multimodal subspace (weight of regression loss $\lambda = 1.0$).

Regression weight λ	0.5	1.0	2.0	4.0	10.0
MNN+Reg	54.04	**55.99**	55.68	55.32	54.12

Table 4: SRN MNN+Reg (VGG_{det}-SPAT1) model's performance (accuracy in %) under different weight λ of regression loss in Eq. 1 (multimodal subspace dimension $m = 128$).

We test different output dimensions of MNN, which is discussed in later section (Table 3). Experiments show that MNN+Reg achieves the best performance among models taking single query phrase as language input, with 8.29% improvement compared to GroundeR [27], and 5.1% to state-of-the-art approach [23].

CRN model. We finetune CRN based on the MNN+Reg SRN, and take VGG_{det}-SPAT1 as input. In training and testing stage, we treat other query phrases from the same caption of the input query phrase as context. We set the input number of context phrases for CRN to be 1 ($M = 1$) and the weight of joint prediction loss $\mu = 1.0$. There is slight improvement (0.32%) increase in accuracy.

MSRC System. For MSRC model's prediction, we fuse CRN's probability as well as SRN's probability, select the proposal with maximum probability and then regress the proposal according to the regression parameters predicted by SRN. We set the weights $\tau = 1.0$ in Eq. (9). The MSRC Full model achieves best performance on Flickr30K Entities (57.53%), with 6.64% increase compared to [23].

Since Flickr30K Entities provides the phrase type for each query, we further compare the detailed phrase localization results. In Table 2, we observe similar boosts in performance by adopting SRN, CRN and fusion by MSRC as in Table 1. However, different models' strengths are different. CCA embedding [23] model is strong in localizing "instruments" while GroundeR [27] is better in localizing "scenes". By using SRN, we observe that the regression network achieves increase in accuracy compared to GroundeR model (VGG_{det}). Typically, there is a large increase in performance of localizing "animals" and "body parts" (with increase of 24% and 13% respectively). By using CRN, we observe that the increase in "scene" is the largest. In the final fusion stage, MSRC Full model achieves more than 6.5%, 5.88%, 2.9% increase in accuracy in all categories (except "instrument" for CCA embedding [23]) compared to GroundeR [27], Wang *et al.* [30] and CCA embedding [23] respectively.

Dimension of multimodal subspace in SRN. To find the relation between SRN's performance and multimodal subspace's dimension, we train and test SRN (MNN+Reg) in five different multimodal subspace dimensions, which are $m = 64, 128, 256, 512, 1024$. The performances are recorded in Table 3. From the results, we observe SRN has lower performance when multimodal subspace has a smaller dimension, which may be caused by lack of trainable parameters to exhibit the model's expressive power. When multimodal subspace has larger dimensions, the performance fluctuates in a small scale, which reflects that SRN is insensitive to multimodal subspace's dimension when m is large.

Weight of regression loss. During training, SRN's loss in Eq. 1 is classification loss plus a weighted regression loss. We test different weights, with $\lambda = 0.5, 1.0, 2.0, 4.0$ to combine regression loss with classification loss. The results are shown in Table 4. From the results, we observe that the hyper parameter λ does not have a big influence on SRN if L^s_{cls} and L^s_{reg} are in the similar range. When λ is large, SRN loses useful information contained in the classification part which helps SRN choose a good proposal to regress. Thus, when $\lambda = 10.0$, we observe a decrease in performance.

4.4 Performance on Refer-it Game

SRN only. Based on Refer-it Game dataset's structure, there is no context information labeled, because there are no image captions for each image. Hence, we only evaluate SRN's performance on Refer-it Game dataset. In testing stage, we choose the proposal with maximum probability predicted only by SRN. This is equal to setting $\tau = 0$ in Eq. (9).

Comparison of different approaches on Refer-it Game dataset is shown in Table 5. From the results, we observe MCB does not have good performance as MNN. This is likely due to MCB needs to maintain a large output dimension to exhibit the model's expressive power. Refer-it Game does not have as much data as Flickr30K Entities. Thus, the training procedure may overfit in early stage. After using regression network, SRN with MCB (MCB+Reg) has comparable performance with GroundeR [27]. By using MNN (MNN+Reg),

Approach	Accuracy (%)
Compared approaches	
SCRC [13]	17.93
GroundeR (VGG$_{\text{cls}}$-SPAT2) [27]	26.93
Spatial Regression models	
MCB+Reg (VGG$_{\text{cls}}$-SPAT2)	26.54
MNN+Reg (VGG$_{\text{cls}}$-SPAT2)	**32.21**

Table 5: Different models’ performance on Refer-it Game. Since there is no context information annotated, we only evaluate SRN models

Multimodal dimension m	64	128	256	512
MNN+Reg	30.72	**32.21**	30.89	31.65

Table 6: SRN MNN+Reg (VGG$_{\text{cls}}$-SPAT2) model’s performance (accuracy in %) under different dimension of multimodal subspace on Refer-it Game dataset. We fix weight of regression loss $\lambda = 1.0$.

Regression weight λ	0.5	1.0	2.0	4.0	10.0
MNN+Reg	31.65	**32.21**	31.56	31.69	32.04

Table 7: SRN MNN+Reg (VGG$_{\text{cls}}$-SPAT2) model’s performance (accuracy in %) under different co-efficient λ of regression loss in Eq. 1. We fix multimodal subspace dimension $m = 128$.

SRN achieves the highest performance, with 5.28% improvement compared to the state-of-the-art method.

Dimension of multimodal subspace in SRN. Similar to Flickr30K Entities, we train and test SRN (MNN+Reg) in four different multimodal subspace dimensions, which are $m = 64, 128, 256, 512$. The performances are recorded in Table 6. From the results, we observe SRN has some fluctuation in accuracy. Overall it is insensitive to multimodal subspace dimension m, which is similar to Table 3.

Weight of regression loss. We test different values of λ in Eq. (1). The results are shown in Table 7. From the results, we find that when λ is small, SRN’s performance is low, because classification part is mostly involved in training. When λ becomes large, SRN’s performance fluctuates and achieves the best at $\lambda = 1.0$, which is similar to results in Table 4.

4.5 Qualitative results

We visualize some phrase grounding results of Flickr30K Entities and Refer-it Game datasets (Fig. 3). For Flickr30K Entities, we show an image and its caption as well as the queries in the caption. For each query, we visualize the ground truth box, the selected proposal box by MSRC system and the regressed bounding box based on the regression parameters predicted by SRN. Since there is no context information in Refer-it Game, we visualize query and its ground truth, with selected proposal and regressed box predicted by SRN.

We observe MSRC achieves good performance in grounding people and clothing, which is consistent with results in Table 2. However, when the query is not clear with no context information, MSRC may ground reasonably incorrect objects (Fig. 3).

5 CONCLUSION

We proposed a novel Multimodal Spatial Regression with semantic Context (MSRC) system, which focuses on phrase grounding problem. Given a query and query-related context information, MSRC system applies a Spatial Regression Network (SRN) to predict the mentioned object’s location based on the proposal bounding box with the highest probability. Besides, MSRC system applies a Context Refinement Network (CRN) to refine the results by encoding context information and adopting a novel joint prediction loss during training stage. MSRC system not only relieves the performance limitation brought from proposal generation system, but also takes advantage of context information to filter out confusing candidates. MSRC system has a significant improvement in performance compared to state-of-the-arts, with 6.64% and 5.28% increase in accuracy on Flickr30K Entities [23] and Refer-it Game [15] datasets respectively.

6 ACKNOWLEDGEMENTS

This paper is based, in part, on research sponsored by the Air Force Research Laboratory and the Defense Advanced Research Projects Agency under agreement number FA8750-16-2-0204. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory and the Defense Advanced Research Projects Agency or the U.S. Government.

REFERENCES

[1] K. Andrej and F.-F. Li. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C Lawrence Z., and D. Parikh. 2015. VQA: Visual question answering. In *ICCV*.

[3] K. Chen, J. Wang, L.-C. Chen, H. Gao, W. Xu, and R. Nevatia. 2015. ABC-CNN: An attention based convolutional neural network for visual question answering. *arXiv preprint arXiv:1511.05960* (2015).

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F Li. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

[5] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2010. The PASCAL Visual Object Classes Challenge. In *IJCV*.

[6] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C Platt, and others. 2015. From captions to visual concepts and back. In *CVPR*.

[7] A. Fukui, D. H Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *EMNLP* (2016).

[8] R. Girshick. 2015. Fast R-CNN. In *ICCV*.

[9] X. Glorot and Y. Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks.. In *Aistats*.

[10] A. Gordo, J. Almazán, J. Revaud, and D. Larlus. 2016. Deep image retrieval: Learning global representations for image search. In *ECCV*.

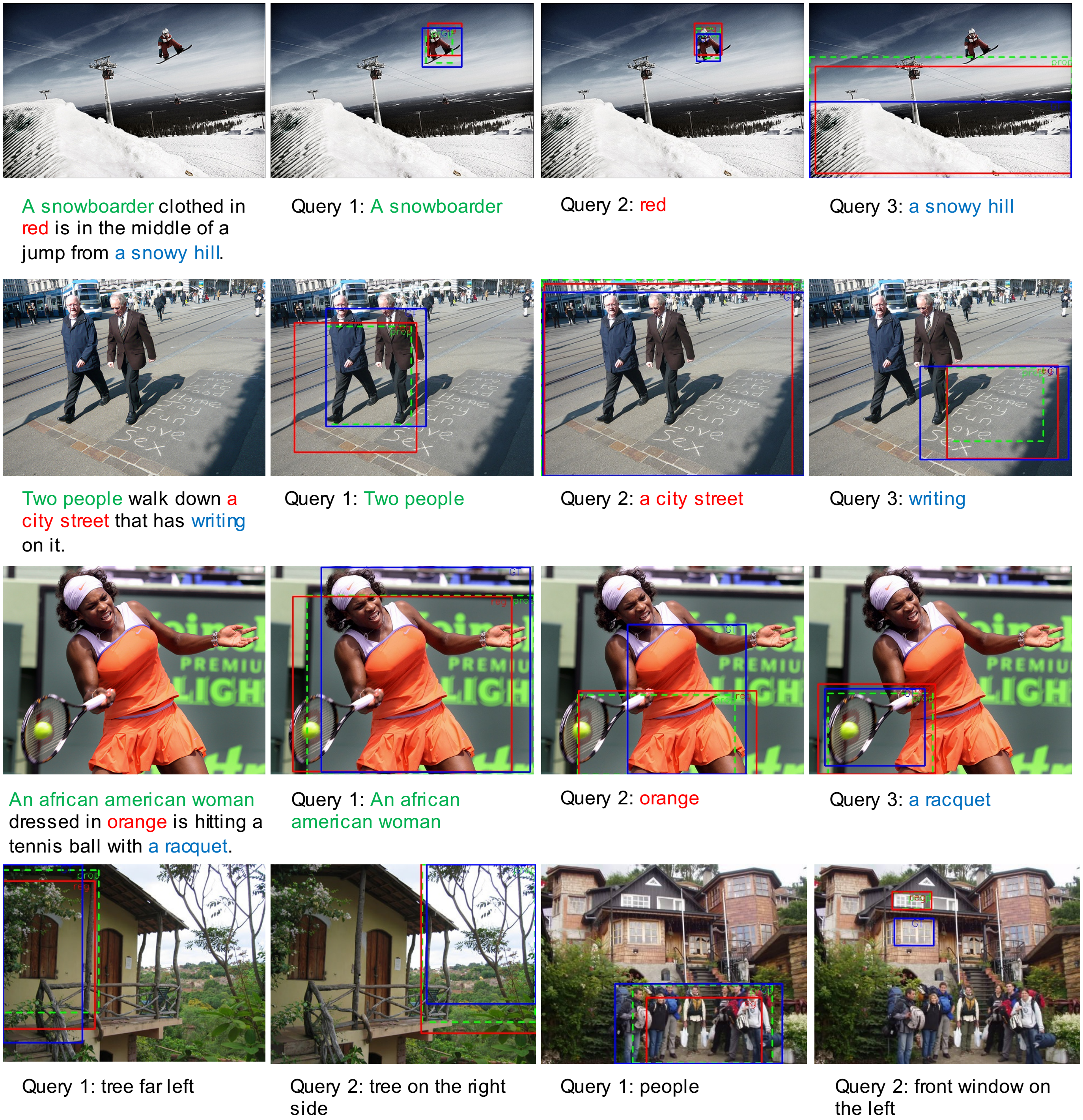
[11] K. He, X. Zhang, S. Ren, and J. Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *CVPR*.

[12] S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural computation* (1997).

[13] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell. 2016. Natural language object retrieval. In *CVPR*.

[14] J. Justin, K. Andrej, and F.-F. Li. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*.

[15] Sahar K., Vicente O., Mark M., and Tamara L. B. 2014. ReferIt Game: Referring to Objects in Photographs of Natural Scenes. In *EMNLP*.



A snowboarder clothed in red is in the middle of a jump from a snowy hill.

Query 1: A snowboarder

Query 2: red

Query 3: a snowy hill

Two people walk down a city street that has writing on it.

Query 1: Two people

Query 2: a city street

Query 3: writing

An african american woman dressed in orange is hitting a tennis ball with a racquet.

Query 1: An african american woman

Query 2: orange

Query 3: a racquet

Query 1: tree far left

Query 2: tree on the right side

Query 1: people

Query 2: front window on the left

Figure 3: Some phrase grounding results generated by MSRC system in Flickr30K and Refer-it Game datasets. We visualize ground truth bounding box, selected proposal box and regressed bounding box in blue, green and red respectively. First three rows are phrase grounding results in Flickr30K Entities dataset. First column is input image and query phrases coming from the same image caption. The $2^{nd} - 4^{th}$ columns correspond to different queries and grounding results. Forth row contains grounding results in Refer-it Game dataset. For different queries, MSRC system is able to localize objects in same images. However, when query is not clear without further context information, MSRC system may ground wrong objects (image in row four, column four).

[16] V. Kantorov, M. Oquab, M. Cho, and I. Laptev. 2016. Contextlocnet: Context-aware deep network models for weakly supervised localization. In *ECCV*.

[17] A. Karpathy, A. Joulin, and F.-F. Li. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *NIPS*.

[18] D. Kingma and J. Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[19] J. Krishnamurthy and T. Kollar. 2013. Jointly learning to parse and perceive: Connecting natural language to the physical world. *TACL* (2013).

[20] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. 2016. SSD: Single shot multibox detector. In *ECCV*.

[21] C. Matuszek, N. FitzGerald, L. Zettlemoyer, L. Bo, and D. Fox. 2012. A joint model of language and perception for grounded attribute learning. *ICML* (2012).

[22] V. K Nagaraja, V. I Morariu, and L. S Davis. 2016. Modeling context between objects for referring expression understanding. In *ECCV*.

[23] B. A Plummer, L. Wang, C. M Cervantes, J. C Caicedo, J. Hockenmaier, and S. Lazebnik. 2016. Flickr30k Entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *IJCV*.

[24] F. Radenović, G. Tolias, and O. Chum. 2016. CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples. In *ECCV*.

[25] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. 2016. You only look once: Unified, real-time object detection. In *CVPR*.

[26] S. Ren, K. He, R. Girshick, and J. Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*.

[27] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. 2016. Grounding of textual phrases in images by reconstruction. In *ECCV*.

[28] K. Simonyan and A. Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* (2014).

[29] J. R. Uijlings, Koen E. Van D. S., T. Gevers, and A. W. Smeulders. 2013. Selective search for object recognition. *IJCV* (2013).

[30] M. Wang, M. Azab, N. Kojima, R. Mihalcea, and J. Deng. 2016. Structured matching for phrase localization. In *ECCV*.

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

[32] C. L. Zitnick and P. Dollár. 2014. Edge boxes: Locating object proposals from edges. In *ECCV*.