Visual Grounding

Luke Ye

52194506006

Outline

- Visual Grounding
  - Background
  - Methods
- Weakly Supervised Visual Grounding
  - Background
  - Methods
- Further Work

Background

Task

- 检索式：预测时给出候选Region
- 回归式：回归Region Box和ground truth比对

评估：
Accuracy （IoU>0.5）

Phrases Grounding

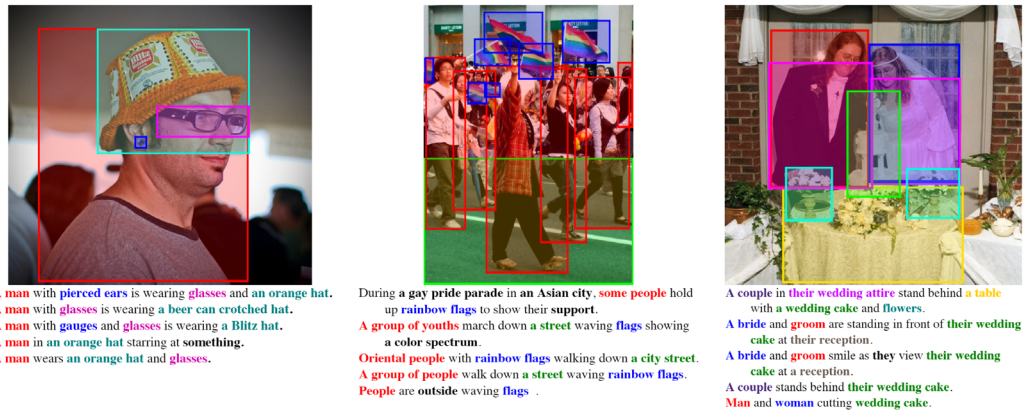
A man with pierced ears is wearing glasses and an orange hat.

Refer expression Grounding

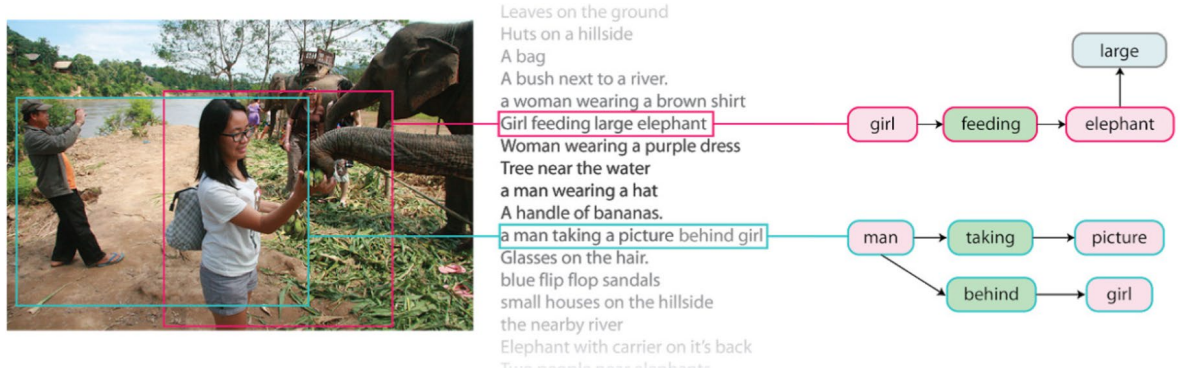
person all the way to the right

Datasets

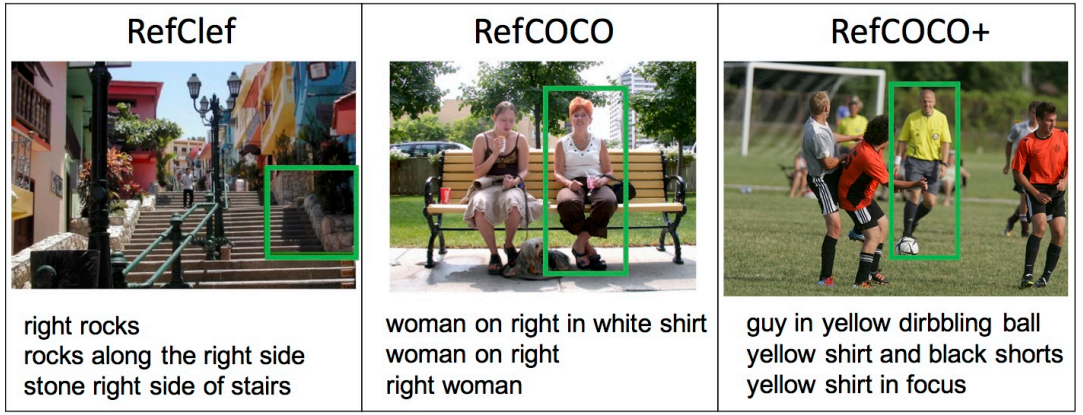
- Flickr30k Entities



- Visual Genome

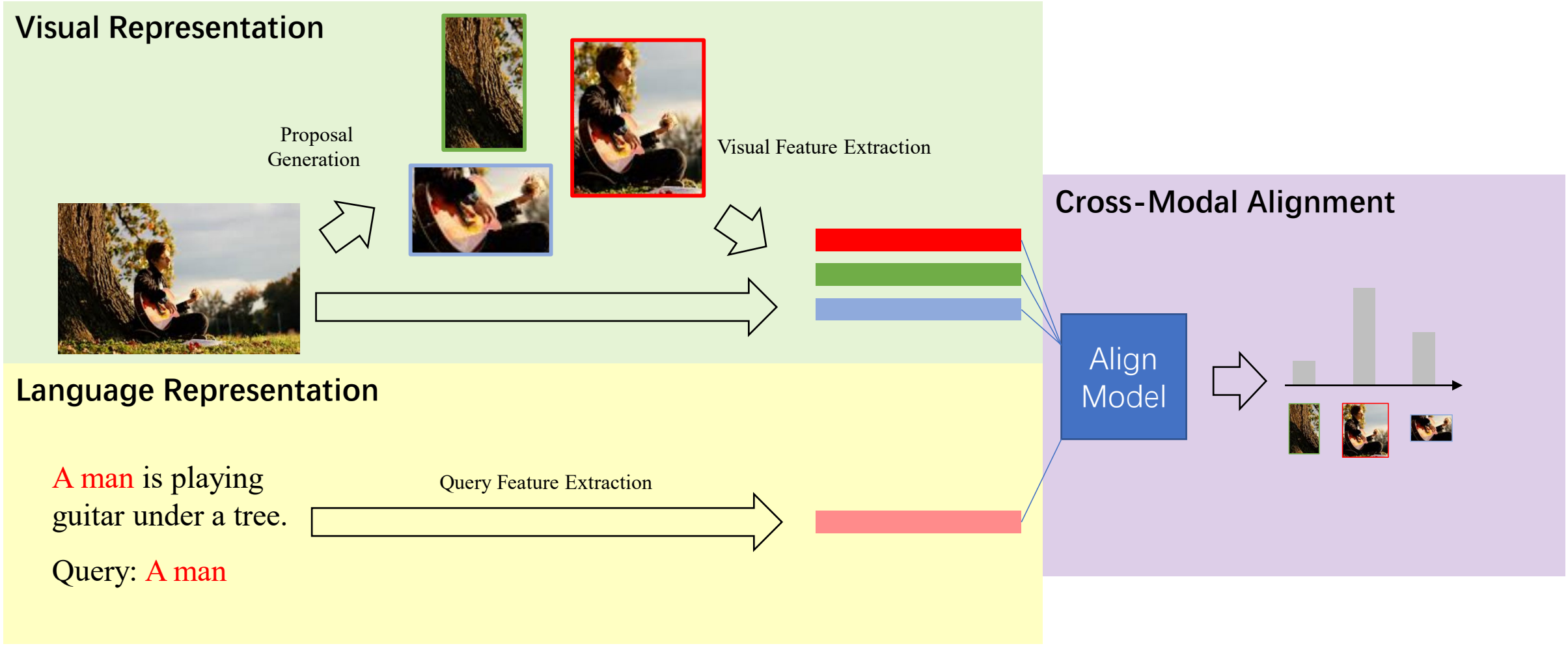


- Refer dataset
  - RefCOCOg
  - RefCOCO/RefCOCO+/RefClef



Methods

Basic Pipeline



ICCV 2015, Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models

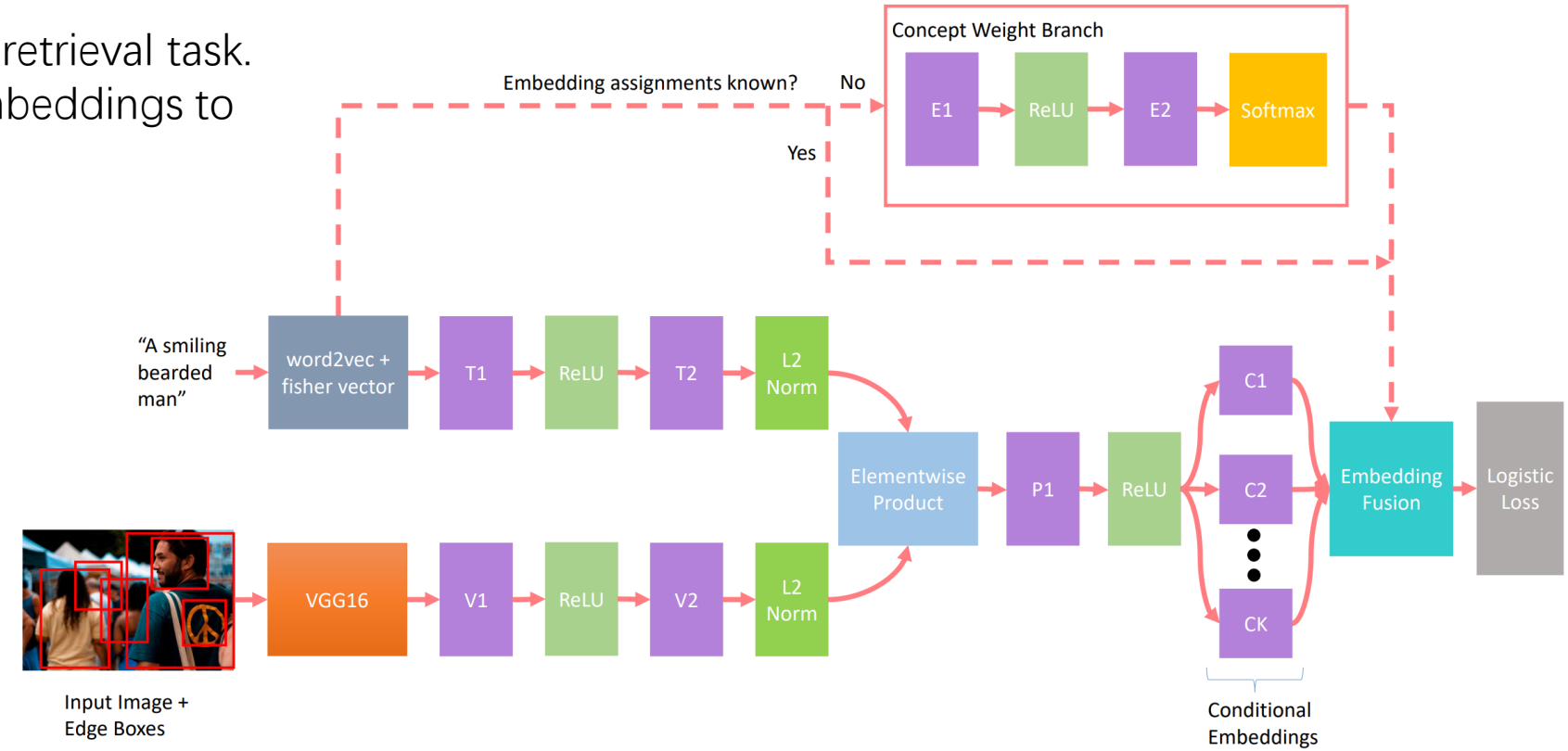
Methods

Retrieval Model with Conditional Embedding

- Model grounding task as a retrieval task.
- Incorporate Conditional Embeddings to better represent concepts.

Embedding Assignment

- Coarse categories
- Nearest cluster center
- Concept weight branch



ECCV 2017, Conditional Image-Text Embedding Networks

Methods

Retrieval Model with Conditional Embedding

Table 1. Phrase localization performance on the Flickr30k Entities test set. (a) State-of-the-art results when predicting a single phrase at a time taken from published works. (b,c) Our baselines and variants using PASCAL-tuned features. (d) Results using Flickr30k-tuned features

Method	Accuracy
(a) **Single Phrase Methods (PASCAL-tuned Features)**[2]	
NonlinearSP [31]	43.89
GroundeR [28]	47.81
MCB [7]	48.69
RtP [25]	50.89
Similarity Network [32]	51.05
IGOP [35]	53.97
SPC [24]	55.49
MCB + Reg + Spatial [3]	51.01
MNN + Reg + Spatial [3]	55.99
(b) **Our Implementation**	
Similarity Network	53.45
Similarity Network + Spatial	54.52
(c) **Conditional Models + Spatial**	
Individual Coarse Category Similarity Networks, $K = 8$	55.32
Individual K-means Similarity Networks, $K = 8$	54.95
CITE, Coarse Categories, $K = 8$	55.42
CITE, Random, $K = 16$	57.58
CITE, K-means, $K = 16$	57.89
CITE, Learned, $K = 4$	58.69
CITE, Learned, $K = 4$, 500 Edge Boxes	59.27
(d) **Flickr30K-tuned Features + Spatial**	
PGN + QRN [4]	60.21
CITE, Learned, $K = 4$, 500 Edge Boxes	**61.89**

Table 6. Localization performance on the ReferIt Game test set. (a) Published results and our Similarity Network baseline. (b) Our best-performing conditional models

Method	Accuracy
(a) **State-of-the-art**	
SCRC [10]	17.93
GroundeR + Spatial [28]	26.93
MCB + Reg + Spatial [3]	26.54
CGRE [21]	31.85
MNN + Reg + Spatial [3]	32.21
IGOP [35]	34.70
Similarity Network + Spatial	31.26
(b) **Conditional Models + Spatial**	
CITE, K-Means, $K = 2$	34.01
CITE, Learned, $K = 12$	34.13

Table 7. Phrase localization performance on Visual Genome. (a) Published results and our Similarity Network baselines. APP refers to ambiguous phrase pruning (see [37] for details). (b) Our best-performing conditional models

Method	Accuracy
(a) **State-of-the-art**	
Densecap [13]	10.1
SCRC [10]	11.0
DBNet [37]	17.5
DBNet (with APP) [37]	21.2
DBNet (with APP, V. Genome-tuned Features) [37]	23.7
Similarity Network	19.76
Similarity Network + Spatial	20.08
(b) **Conditional Models + Spatial**	
CITE, K-Means, $K = 12$	23.67
CITE, Learned, $K = 12$	24.43

ECCV 2017, Conditional Image-Text Embedding Networks

Methods

Spatial Regression with Semantic Context

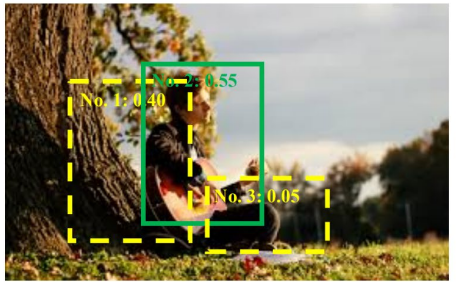
Existing methods' limitations:

- Proposal generation system fails to provide good proposals.
- Phrases are considered as unrelated to each other

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

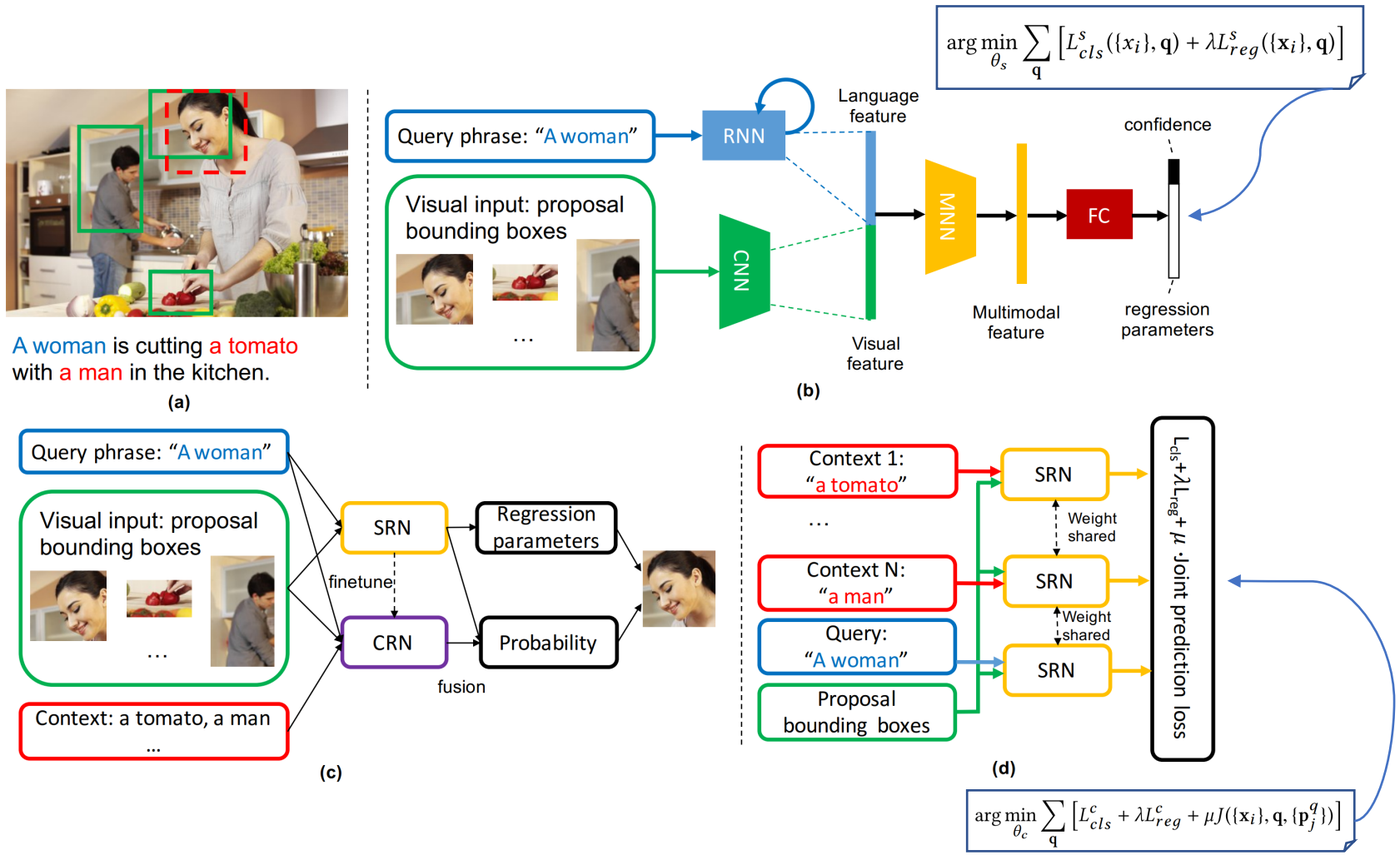
MM 2017, MSRC: Multimodal Spatial Regression with Semantic Context for Phrase Grounding

Methods

Spatial Regression with Semantic Context

(a) Example input
(b) Regression model (SRN)
(d) Context penalizing (CRN)
(c) Grounding model (MSRC)

1. Training SRN
2. Finetuning CRN by the pre-trained SRN



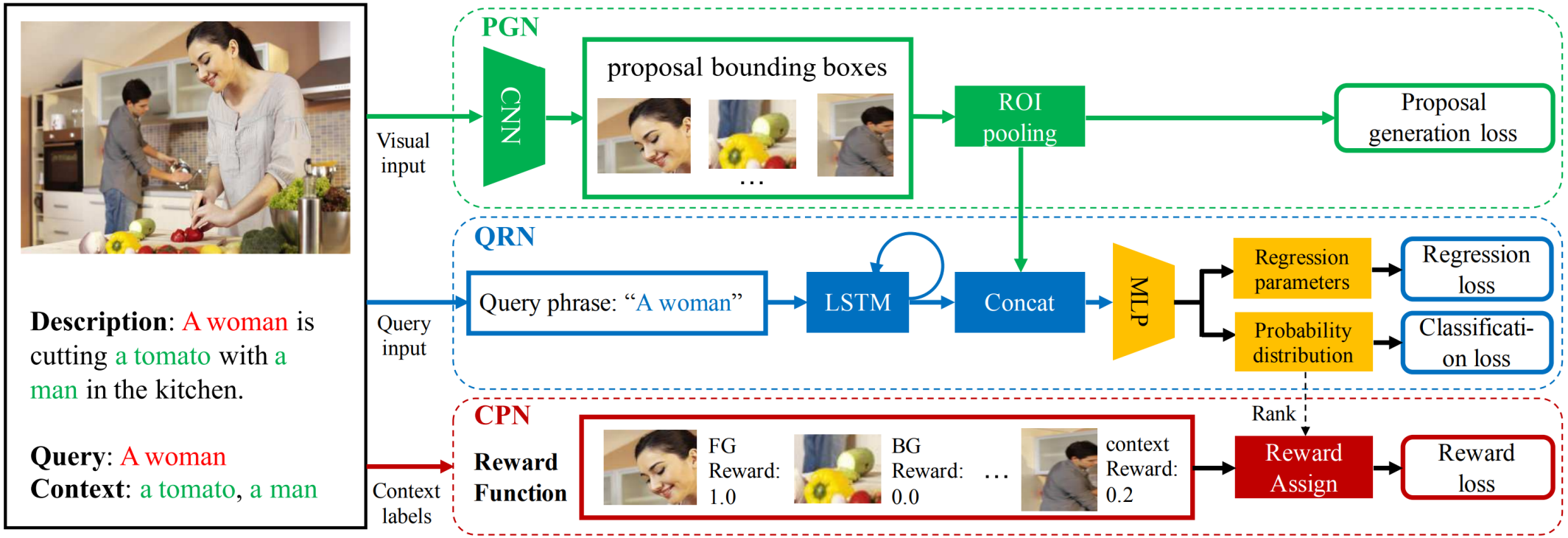
MM 2017, MSRC: Multimodal Spatial Regression with Semantic Context for Phrase Grounding

Spatial Regression with Semantic Context

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**
CITE, Learned, K = 4 + Spatial	75.95	58.50	30.78	77.03	79.25	48.15	58.78	43.24

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

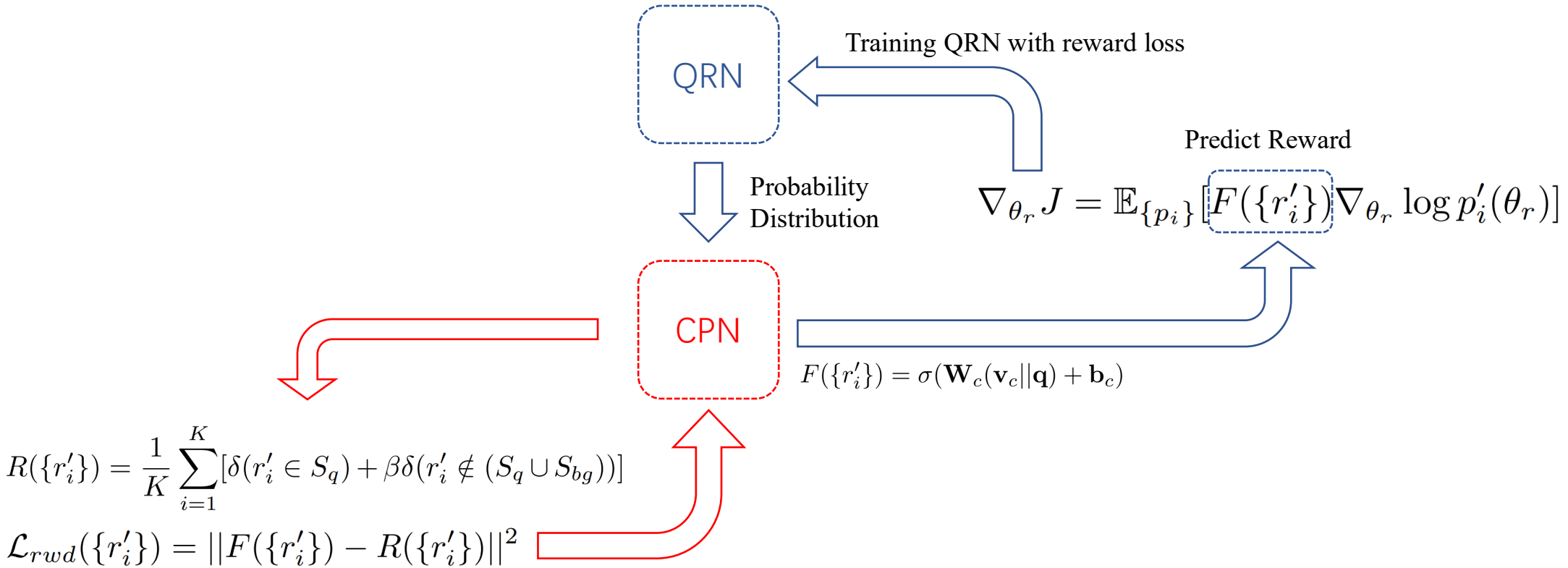
Utilize Context Information through RL



ICCV 2017, Query-Guided Regression Network with Context Policy for Phrase Grounding

Methods

Utilize Context Information through RL



ICCV 2017, Query-Guided Regression Network with Context Policy for Phrase Grounding

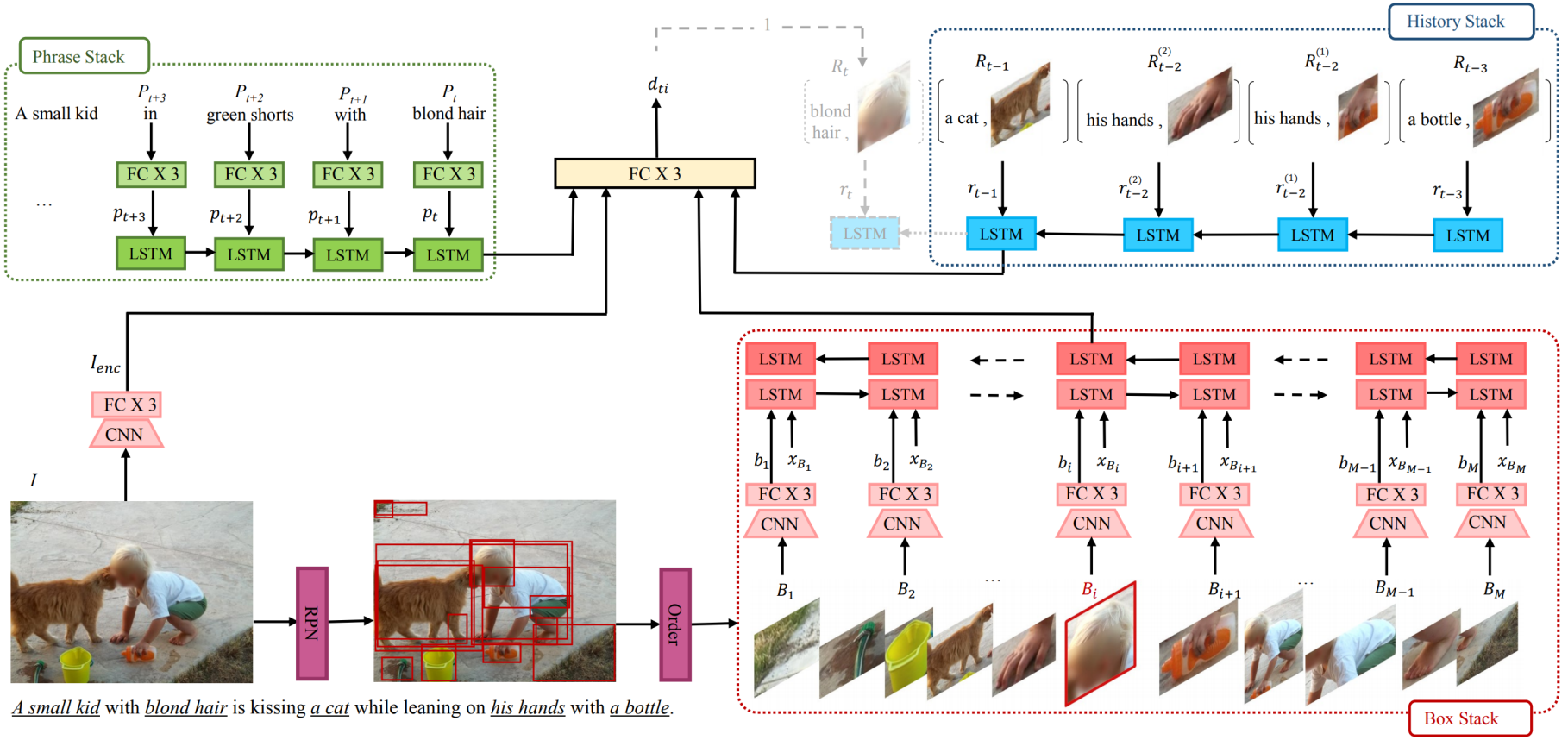
Utilize Context Information through RL

Phrase Type	people	clothing	body parts	animals	vehicles	instruments	scene	other
GroundeR (VGG_{cls}) [29]	53.80	34.04	7.27	49.23	58.75	22.84	52.07	24.13
GroundeR (VGG_{det}) [29]	61.00	38.12	10.33	62.55	68.75	36.42	58.18	29.08
Structured Matching [34]	57.89	34.61	15.87	55.98	52.25	23.46	34.22	26.23
CCA embedding [25]	64.73	46.88	17.21	65.83	68.75	37.65	51.39	31.77
SS+QRN	68.24	47.98	20.11	73.94	73.66	29.34	66.00	38.32
PGN+QRN	75.08	55.90	20.27	73.36	68.95	45.68	65.27	38.80
QRC Net	**76.32**	**59.58**	**25.24**	**80.50**	**78.25**	**50.62**	**67.12**	**43.60**

Table 6. Phrase grounding performances for different phrase types defined in Flickr30K Entities. Accuracy is in percentage.

Methods

Reasoning on Context Information



CVPR 2019, Neural Sequential Phrase Grounding (SeqGROUND)

Reasoning on Context Information

Method	Accuracy
SMPL [48]	42.08
NonlinearSP [47]	43.89
GroundeR [37]	47.81
MCB [11]	48.69
RtP [34]	50.89
Similarity Network [46]	51.05
RPN+QRN [3]	53.48
IGOP [55]	53.97
SPC+PPC [33]	55.49
SS+QRN [3]	55.99
CITE [32]	59.27
SeqGROUND	**61.60**

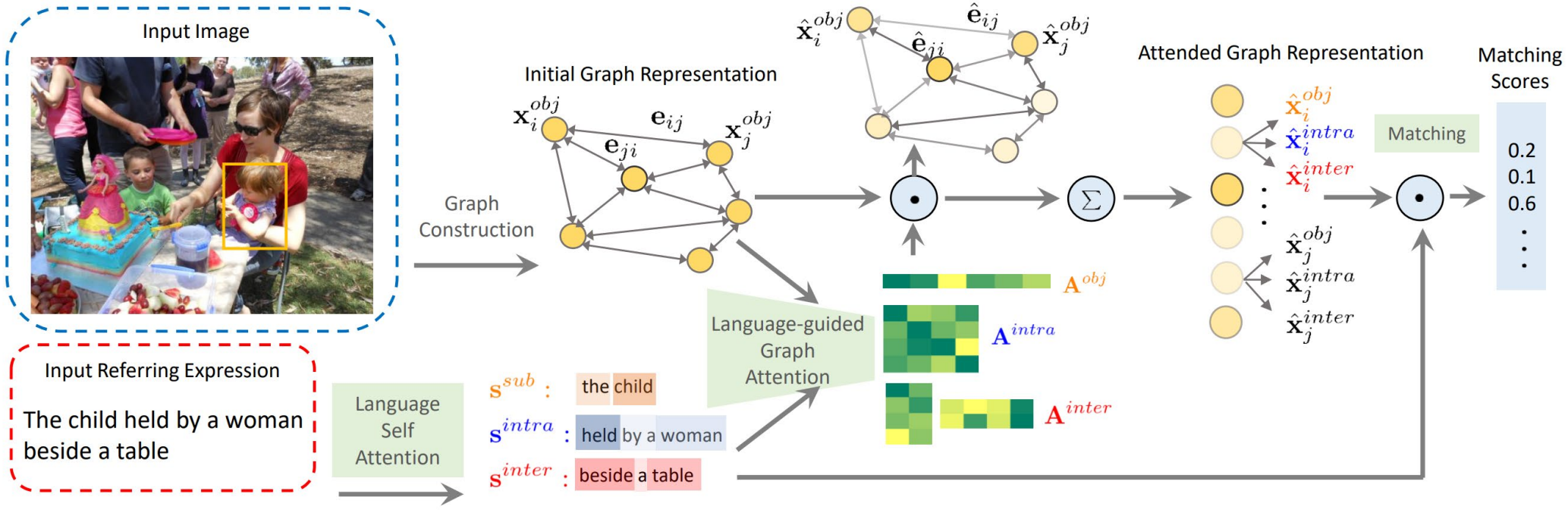
Table 1: **Phrase grounding accuracy** (in percentage) of the state-of-the-art methods on the Flickr30k Entities dataset.

Method	people	clothing	body parts	animals	vehicles	instruments	scene	other
SMPL [48]	57.89	34.61	15.87	55.98	52.25	23.46	34.22	26.23
GroundeR [37]	61.00	38.12	10.33	62.55	68.75	36.42	58.18	29.08
RtP [34]	64.73	46.88	17.21	65.83	68.72	37.65	51.39	31.77
IGOP [55]	68.71	56.83	19.50	70.07	73.72	39.50	60.38	32.45
SPC+PPC [33]	71.69	50.95	25.24	76.23	66.50	35.80	51.51	35.98
CITE [32]	73.20	52.34	30.59	76.25	75.75	48.15	55.64	42.83
SeqGROUND	76.02	56.94	26.18	75.56	66.00	39.36	68.69	40.60

Table 2: **Comparison of phrase grounding accuracy** (in percentage) over coarse categories on Flickr30K dataset.

Methods

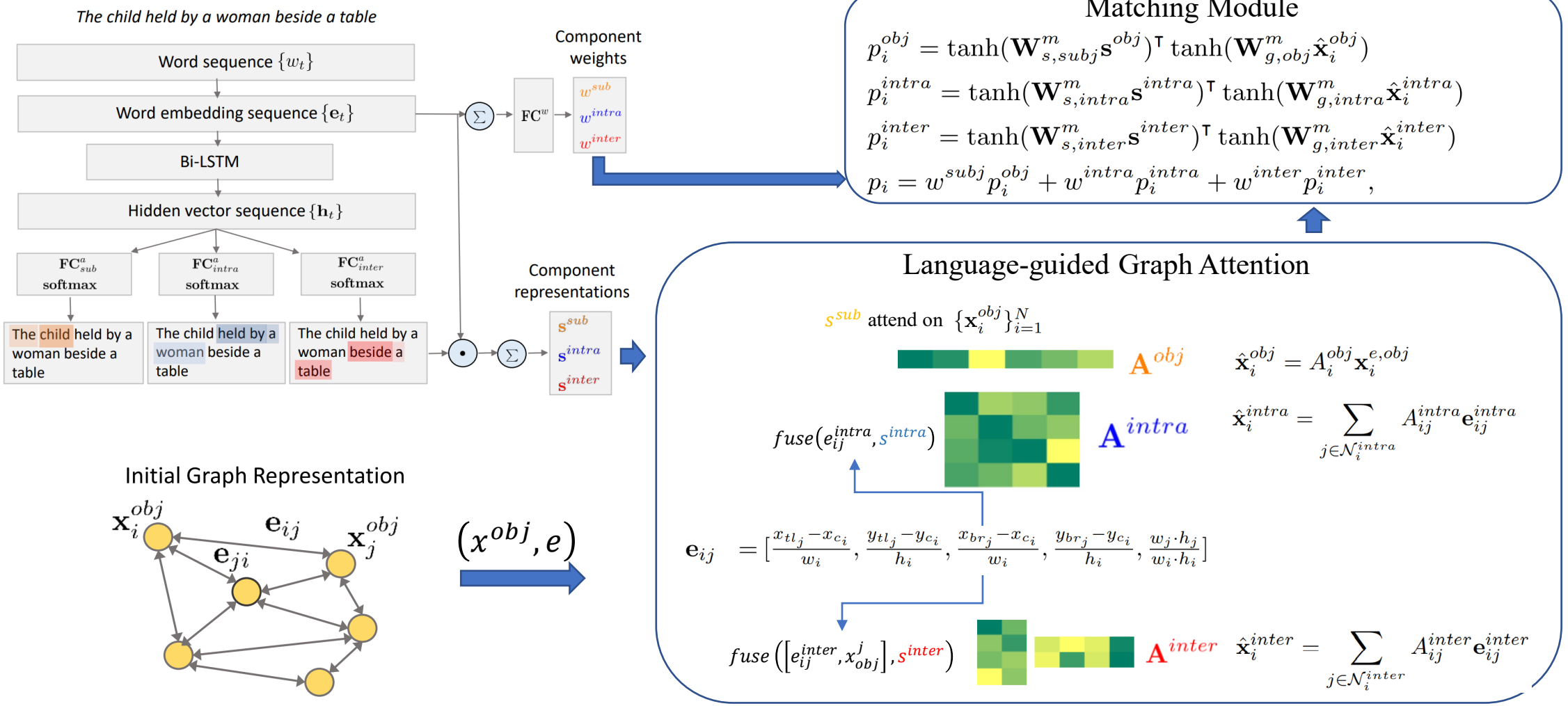
Encode Context Information through Graph



CVPR 2019, Neighbourhood Watch: Referring Expression Comprehension via Language-Guided Graph Attention Networks

Methods

Encode Context Information through Graph



CVPR 2019, Neighbourhood Watch: Referring Expression Comprehension via Language-Guided Graph Attention Networks

Encode Context Information through Graph

Methods	RefCOCO			RefCOCO+			RefCOCOg		
	val	testA	testB	val	testA	testB	val*	val	test
MMI [15]	-	71.72	71.09	-	58.42	51.23	62.14	-	-
visdif [28]	-	67.57	71.19	-	52.44	47.51	59.25	-	-
visdif+MMI [28]	-	73.98	76.59	-	59.17	55.62	64.02	-	-
NegBag [16]	76.90	75.60	78.00	-	-	-	-	-	68.40
CMN [6]	-	75.94	79.57	-	59.29	59.34	69.3	-	-
listener [29]	77.48	76.58	78.94	60.5	61.39	58.11	71.12	69.93	69.03
speaker+listener+reinforcer [29]	78.14	76.91	80.1	61.34	63.34	58.42	72.63	71.65	71.92
speaker+**listener**+reinforcer [29]	78.36	77.97	79.86	61.33	63.1	58.19	72.02	71.32	71.72
VariContxt [30]	-	78.98	82.39	-	62.56	62.90	73.98	-	-
ParallelAttn [32]	81.67	80.81	81.32	64.18	66.31	61.46	69.47	-	-
AccumulateAttn [3]	81.27	81.17	80.01	65.56	**68.76**	60.63	73.18	-	-
MattNet [27]	80.94	79.99	82.3	63.07	65.04	61.77	73.08	73.04	72.79
Ours-LGRANs	**82.0**	**81.2**	**84.0**	**66.6**	67.6	**65.5**	-	**75.4**	**74.7**

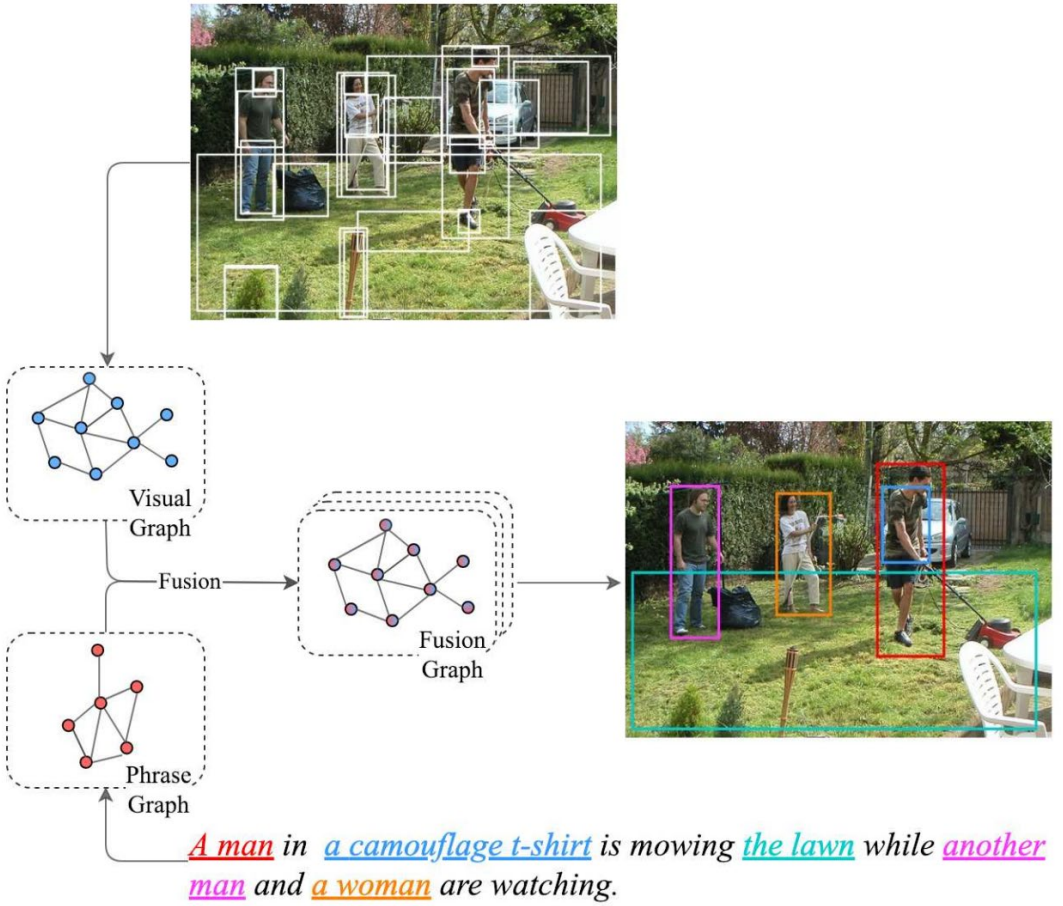
Methods

Encode Context Information through Graph

Existing Graph based methods are not general

- Handcraft feature input.
- Manual designed message passing.

Proposal a general GNN model for any multi-modal assignment problem where some contextual relations between elements in each modality exist.

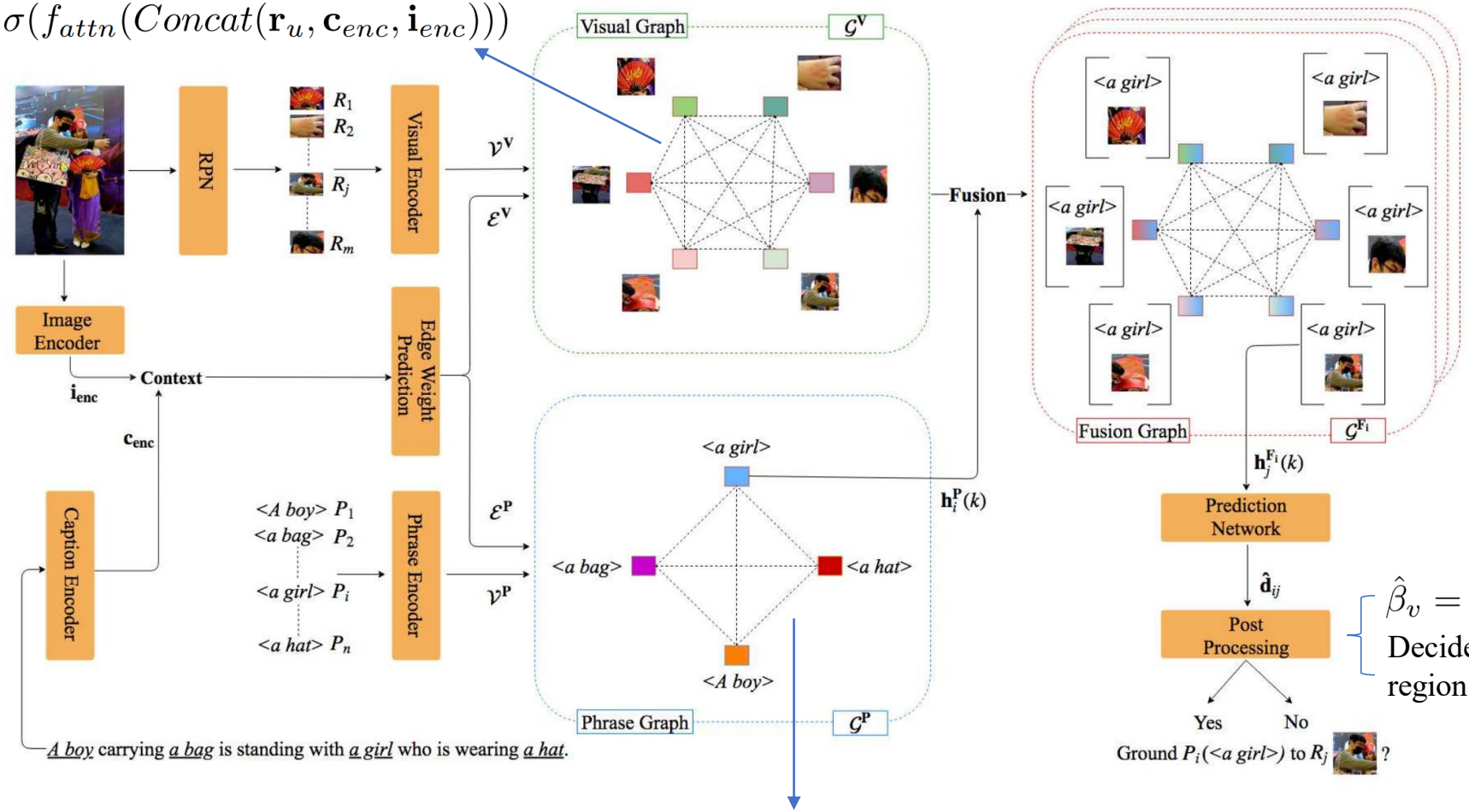


ICCV 2019, G3raphGround: Graph-Based Language Grounding

Methods

Encode Context Information through Graph

$$\alpha_u = \sigma(f_{attn}(Concat(\mathbf{r}_u, \mathbf{c}_{enc}, \mathbf{i}_{enc})))$$



$$\hat{\beta}_v = \sigma(f_{count}(h_v^P(k)))$$

Decide select one or multiple region for each phrase

$$\mathbf{A}^P_{u,v} = \mathbf{A}^P_{v,u} = \sigma(f_{adj}(Concat(\mathbf{p}_u + \mathbf{p}_v, \mathbf{c}_{enc}, \mathbf{i}_{enc})))$$

ICCV 2019, G3raphGround: Graph-Based Language Grounding

Encode Context Information through Graph

Method	Accuracy
SMPL [30]	42.08
NonlinearSP [29]	43.89
GroundeR [26]	47.81
MCB [10]	48.69
RtP [24]	50.89
Similarity Network [28]	51.05
IGOP [38]	53.97
SPC+PPC [23]	55.49
SS+QRN (VGG_{det}) [6]	55.99
CITE [22]	59.27
SeqGROUND [9]	61.60
CITE [22] (finetuned)	61.89
QRC Net [6] (finetuned)	65.14
G^{3}RAPHGROUND++	**66.93**

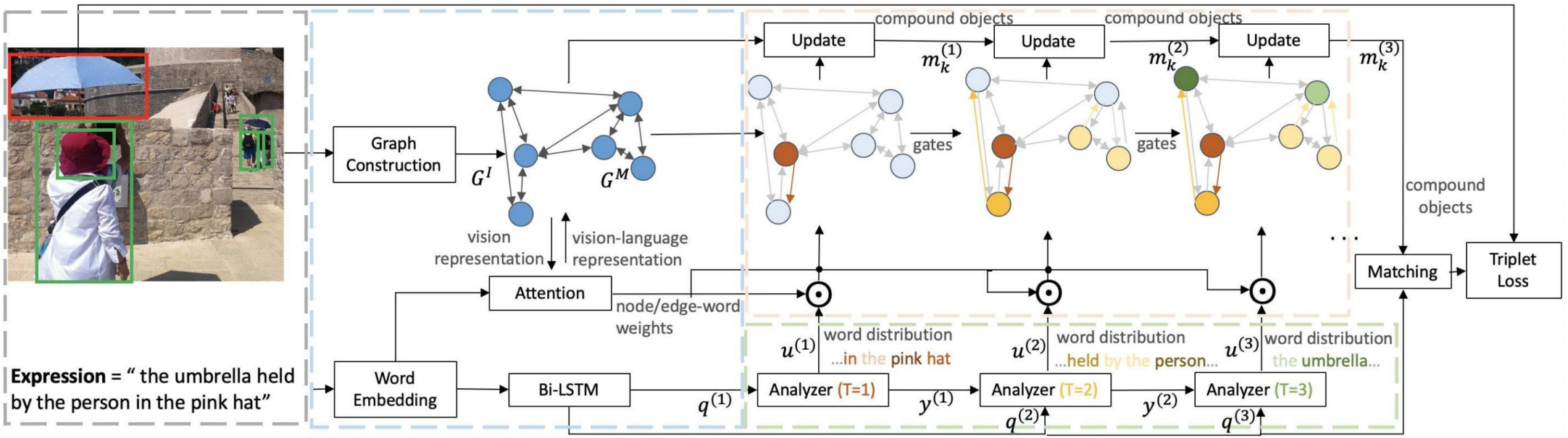
Table 1. **State-of-the-art comparison on Flickr30k.** Phrase grounding accuracy on the test set reported in percentages.

Method	Accuracy
SCRC [12]	17.93
MCB + Reg + Spatial [5]	26.54
GroundeR + Spatial [26]	26.93
Similarity Network + Spatial [28]	31.26
CGRE [20]	31.85
MNN + Reg + Spatial [5]	32.21
EB+QRN (VGG_{cls}-SPAT) [6]	32.21
CITE [22]	34.13
IGOP [38]	34.70
QRC Net [6] (finetuned)	44.07
G^{3}RAPHGROUND++	**44.91**

Table 2. **State-of-the-art comparison on ReferIt Game.** Phrase grounding accuracy on the test set reported in percentages.

Encode Context Information through Graph

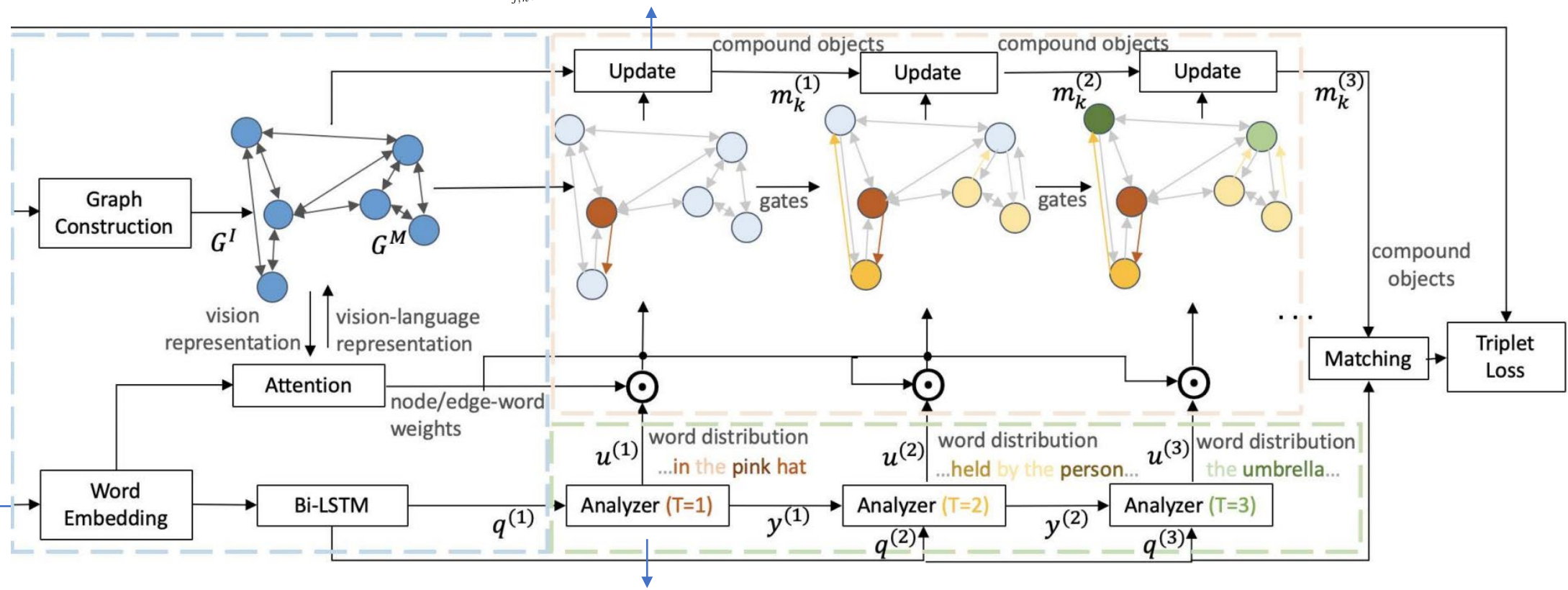
Edge Type:	inside	cover	overlap	No relationship	direction
e_{ij}	1	2	3	0	$4+\left\lfloor\frac{\theta_{ij}+22.5}{45}\right\rfloor$



ICCV 2019, Dynamic Graph Attention for Referring Expression Comprehension

Encode Context Information through Graph

$$\overleftarrow{\boldsymbol{m}}_k^{(t)} = \sum_{e_{j,k}>0} \nu_{e_{j,k}}^{(t)} (\overleftarrow{\boldsymbol{W}} \boldsymbol{m}_j^{(t-1)} p_j^{(t-1)} + \overleftarrow{\boldsymbol{b}}_{e_{j,k}}),$$



- $\alpha_{k,l}$ 词l关于节点k的概率
- β_l 词l关于边类型的概率分布

$r_l^{(t)}$词l是否需要参与视觉推理

ICCV 2019, Dynamic Graph Attention for Referring Expression Comprehension

Encode Context Information through Graph

		RefCOCO			RefCOCO+			RefCOCOg	
	feature	val	testA	testB	val	testA	testB	val	test
MMI [18]	vgg16	-	63.15	64.21	-	48.73	42.13	-	-
Neg Bag [19]	vgg16	76.90	75.60	78.00	-	-	-	-	68.40
CG [16]	vgg16	-	74.04	73.43	-	60.26	55.03	-	-
Attr [13]	vgg16	-	78.85	78.07	-	61.47	57.22	-	-
CMN [7]	vgg16	-	75.94	79.57	-	59.29	59.34	-	-
Speaker [31]	vgg16	76.18	74.39	77.30	58.94	61.29	56.24	-	-
Spearker+**Listener**+Reinforcer[32]	vgg16	78.36	77.97	79.86	61.33	63.10	58.19	71.32	71.72
Speaker+Listener+Reinforcer [32]	vgg16	79.56	78.95	80.22	62.26	64.60	59.62	71.65	71.92
AccumulateAttn [4]	vgg16	81.27	81.17	80.01	65.56	68.76	60.63	-	-
ParallelAttn [33]	vgg16	81.67	80.81	81.32	64.18	66.31	61.46	-	-
MAttNet [30]	vgg16	80.94	79.99	82.30	63.07	65.04	61.77	73.04	72.79
Ours DGA	vgg16	**83.73**	**83.56**	**82.51**	**68.99**	**72.72**	**62.98**	**75.76**	**75.79**
MAttNet [30]	resnet101	85.65	85.26	84.57	71.01	75.13	66.17	78.10	78.12
Ours DGA	resnet101	**86.34**	**86.64**	**84.79**	**73.56**	**78.31**	**68.15**	**80.21**	**80.26**

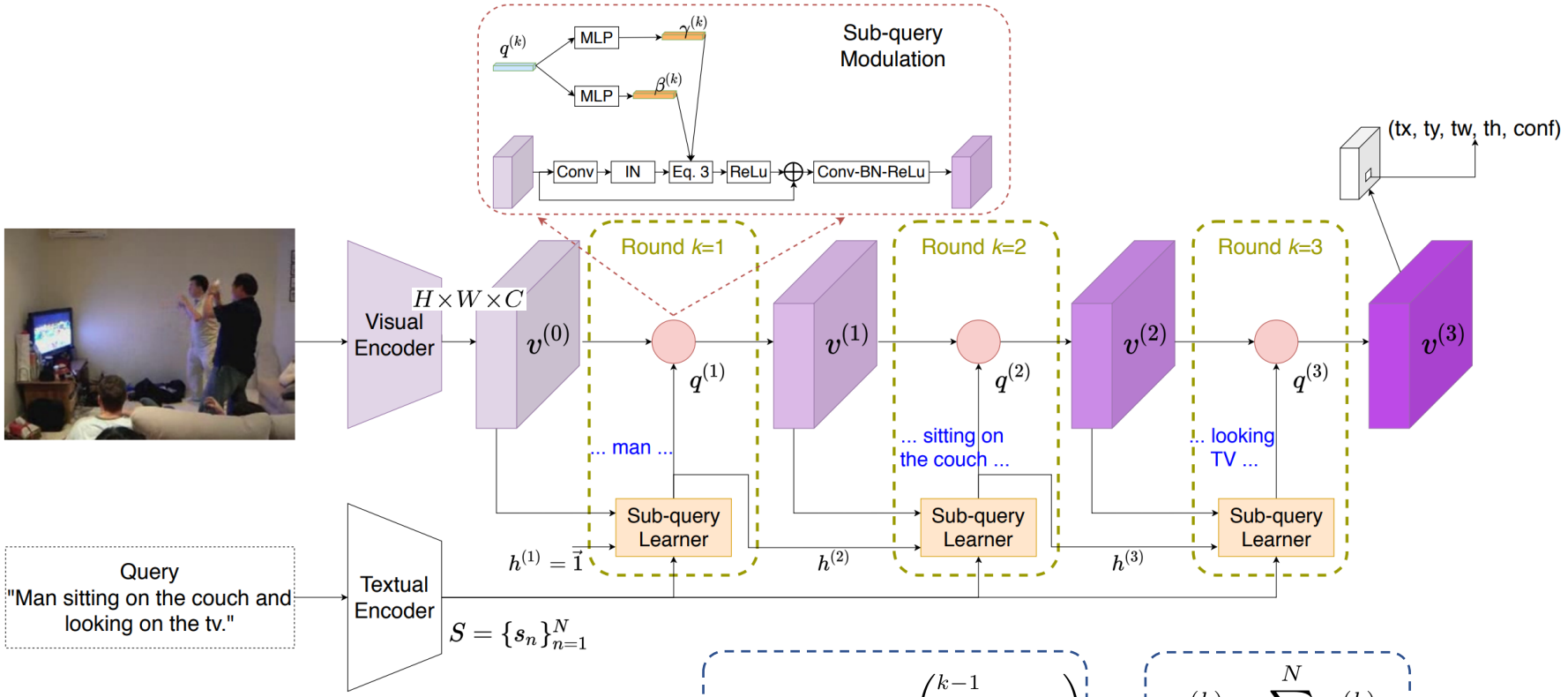
Table 1. Comparison with state-of-the-art methods on RefCOCO, RefCOCO+ and RefCOCOg when ground-truth bounding boxes are used. The best performing method is marked in bold.

	RefCOCO		RefCOCO+		RefCOCOg
	testA	testB	testA	testB	test
MMI [18]	64.90	54.51	54.03	42.81	-
Neg Bag [19]	58.60	56.40	-	-	49.50
CG [16]	67.94	55.18	57.05	43.33	-
Attr [13]	72.08	57.29	57.97	46.20	-
CMN [7]	71.03	65.77	54.32	47.76	-
Speaker [31]	67.64	55.16	55.81	43.43	-
S+**L**+R [32]	72.94	62.98	58.68	47.68	59.63
S+L+R [32]	72.88	63.43	60.43	48.74	59.21
ParallelAttn [33]	75.31	65.52	61.34	50.86	-
Ours DGA	**78.42**	**65.53**	**69.07**	**51.99**	**63.28**

Table 2. Comparison with the state-of-the-art methods on RefCOCO, RefCOCO+ and RefCOCOg when detected objects are used. The best performing method is marked in bold.

Methods

Recursive Sub-query Construction



$$L_{div} = \left\| A^T A \odot (\mathbb{1} - I) \right\|_F^2, \quad L_{cover} = \left\| \vec{1} - \min\left(\sum_{i=1}^{K} \vec{\alpha}^{(i)}, \vec{1} \right) \right\|_1$$

$$\vec{h}^{(k)} = \vec{1} - \min\left(\sum_{i=1}^{k-1} \vec{\alpha}^{(i)}, \vec{1} \right)$$

$$q^{(k)} = \sum_{n=1}^{N} \alpha_n^{(k)} s_n$$

$$\alpha_n^{(k)} = softmax\left[W_{a1}^{(k)} \tanh\left(W_{a0}^{(k)} h_n^{(k)} (\bar{v}^{(k-1)} \odot s_n) + b_{a0}^{(k)} \right) + b_{a1}^{(k)} \right]$$

Recursive Sub-query Construction

Method	Feature	RefCOCO			RefCOCO+			RefCOCOg			Time
		val	testA	testB	val	testA	testB	val-g	val-u	test-u	(ms)
Two-stage Methods											
MMI [29]	VGG16-Imagenet	-	64.90	54.51	-	54.03	42.81	45.85	-	-	-
Neg Bag [30]	VGG16-Imagenet	-	58.60	56.40	-	-	-	-	-	49.50	-
CMN [14]	VGG16-COCO	-	71.03	65.77	-	54.32	47.76	57.47	-	-	-
ParallelAttn [52]	VGG16-Imagenet	-	75.31	65.52	-	61.34	50.86	58.03	-	-	-
VC [51]	VGG16-COCO	-	73.33	67.44	-	58.40	53.18	62.30	-	-	-
LGRAN [42]	VGG16-Imagenet	-	76.6	66.4	-	64.0	53.4	61.78	-	-	-
SLR [50]	Res101-COCO	69.48	73.71	64.96	55.71	60.74	48.80	-	60.21	59.63	-
MAttNet [48]	Res101-COCO	76.40	80.43	69.28	64.93	70.26	56.00	-	66.67	67.01	320
上一篇 DGA [46]	Res101-COCO	-	78.42	65.53	-	69.07	51.99	-	-	63.28	341
One-stage Methods											
SSG [5]	Darknet53-COCO	-	76.51	67.50	-	62.14	49.27	47.47	58.80	-	25
One-Stage-BERT [47]	Darknet53-COCO	72.05	74.81	67.59	55.72	60.37	48.54	48.14	59.03	58.70	23
One-Stage-BERT*	Darknet53-COCO	72.54	74.35	68.50	56.81	60.23	49.60	56.12	61.33	60.36	23
Ours-Base	Darknet53-COCO	76.59	78.22	**73.25**	63.23	66.64	55.53	60.96	64.87	64.87	26
Ours-Large	Darknet53-COCO	**77.63**	**80.45**	72.30	**63.59**	**68.36**	**56.81**	**63.12**	**67.30**	**67.20**	36

Make use of training data

Existing methods didn't make full use of the training data

- neglect multiple textual-visual correspondences

Apply an erasing methods to build more training examples.

Make use of training data

MattNet

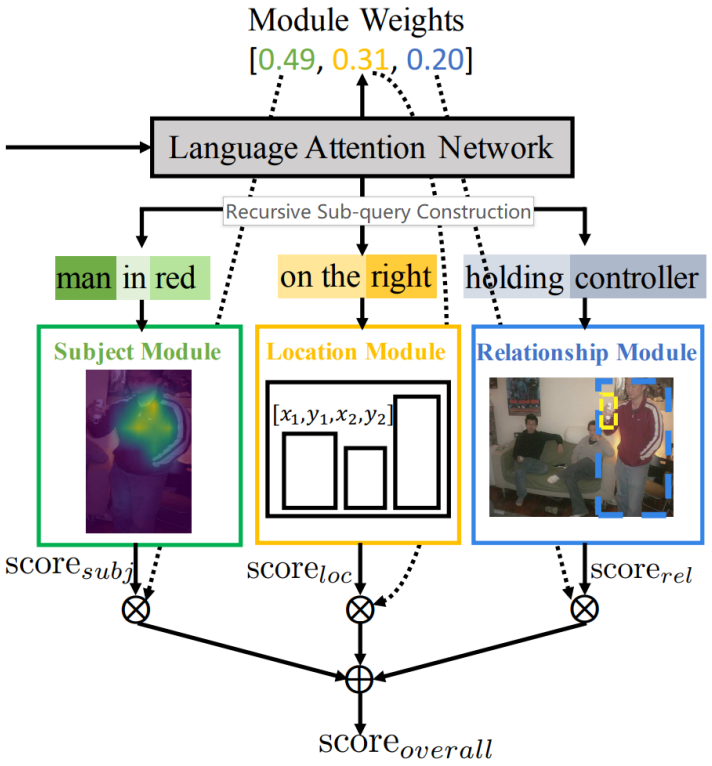
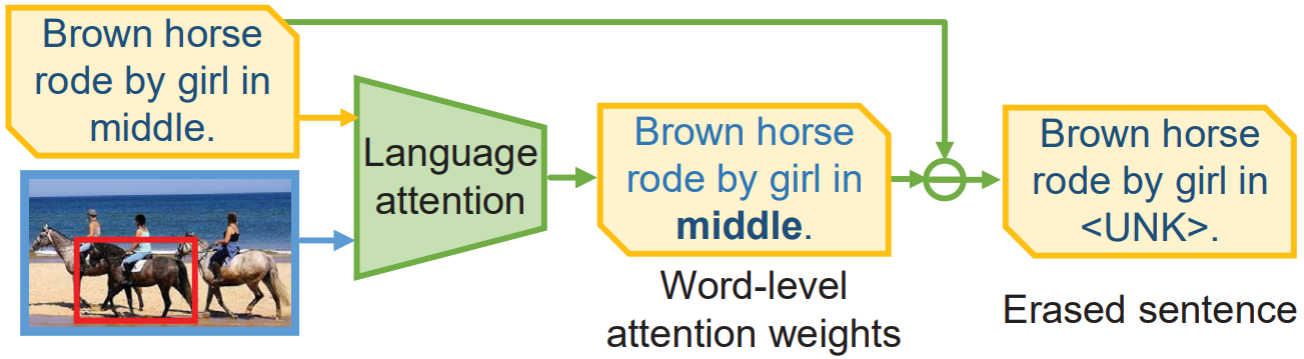
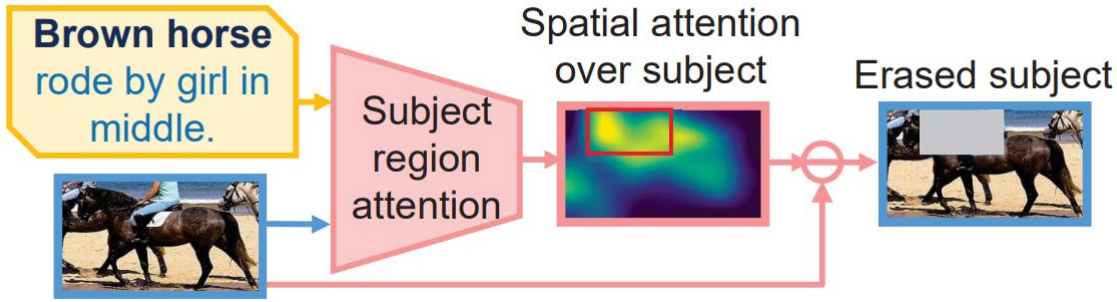


Image-aware query sentence erasing



Sentence-aware subject/context region erasing



CVPR 2018, MAttNet: Modular Attention Network for Referring Expression Comprehension
CVPR 2019, Improving Referring Expression Grounding With Cross-Modal Attention-Guided Erasing

Make use of training data

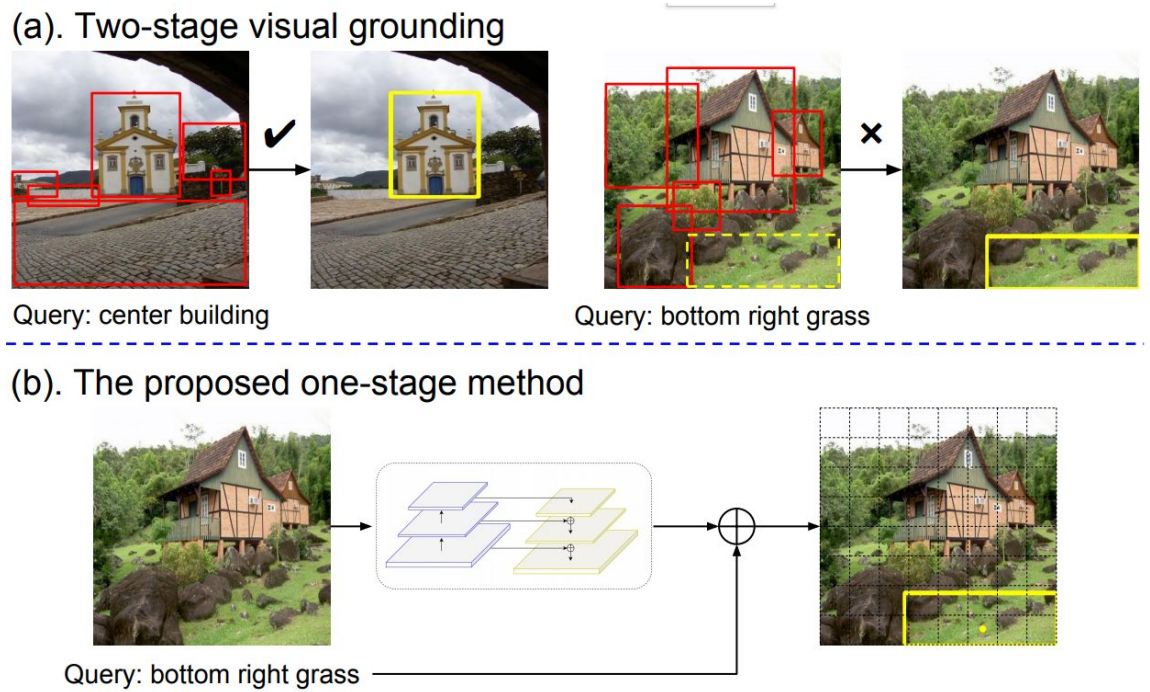
	test setting	RefCOCO			RefCOCO+			RefCOCOg		
		val	testA	testB	val	testA	testB	val*	val	test
MMI [22]	ground-truth	-	71.72	71.09	-	58.42	51.23	62.14	-	-
NegBag [23]	ground-truth	76.90	75.60	78.00	-	-	-	-	-	68.40
visdif+MMI [39]	ground-truth	-	73.98	76.59	-	59.17	55.62	64.02	-	-
Luo *et al*. [21]	ground-truth	-	74.04	73.43	-	60.26	55.03	65.36	-	-
CMN [10]	ground-truth	-	-		-	-	-	69.30	-	-
Speaker/visdif [39]	ground-truth	76.18	74.39	77.30	58.94	61.29	56.24	59.40	-	-
S-L-R [40]	ground-truth	79.56	78.95	80.22	62.26	64.60	59.62	72.63	71.65	71.92
VC [41]	ground-truth	-	78.98	82.39	-	62.56	62.90	73.98	-	-
Attr [17]	ground-truth	-	78.05	78.07	-	61.47	57.22	69.83	-	-
Accu-Att [4]	ground-truth	81.27	81.17	80.01	65.56	68.76	60.63	73.18	-	-
PLAN [43]	ground-truth	81.67	80.81	81.32	64.18	66.31	61.46	69.47	-	-
Multi-hop Film [31]	ground-truth	84.9	87.4	83.1	73.8	**78.7**	65.8	71.5	-	-
MattNet [38]	ground-truth	85.65	85.26	84.57	71.01	75.13	66.17	-	78.10	78.12
CM-Att	ground-truth	86.23	86.57	85.36	72.36	74.64	67.07	-	78.68	78.58
CM-Att-Erase	ground-truth	**87.47**	**88.12**	**86.32**	**73.74**	77.58	**68.85**	-	**80.23**	**80.37**
S-L-R [40]	det proposal	69.48	73.71	64.96	55.71	60.74	48.80	-	60.21	59.63
Luo [21]	det proposal	-	67.94	55.18	-	57.05	43.33	49.07	-	-
PLAN [43]	det proposal	-	75.31	65.52	-	61.34	50.86	58.03	-	-
MattNet [38]	det proposal	76.40	80.43	69.28	64.93	70.26	56.00	-	66.67	67.01
CM-Att	det proposal	76.76	82.16	70.32	66.42	72.58	57.23	-	67.32	67.55
CM-Att-Erase	det proposal	**78.35**	**83.14**	**71.32**	**68.09**	**73.65**	**58.03**	-	**67.99**	**68.67**

CVPR 2019, Improving Referring Expression Grounding With Cross-Modal Attention-Guided Erasing

One-Stage Method

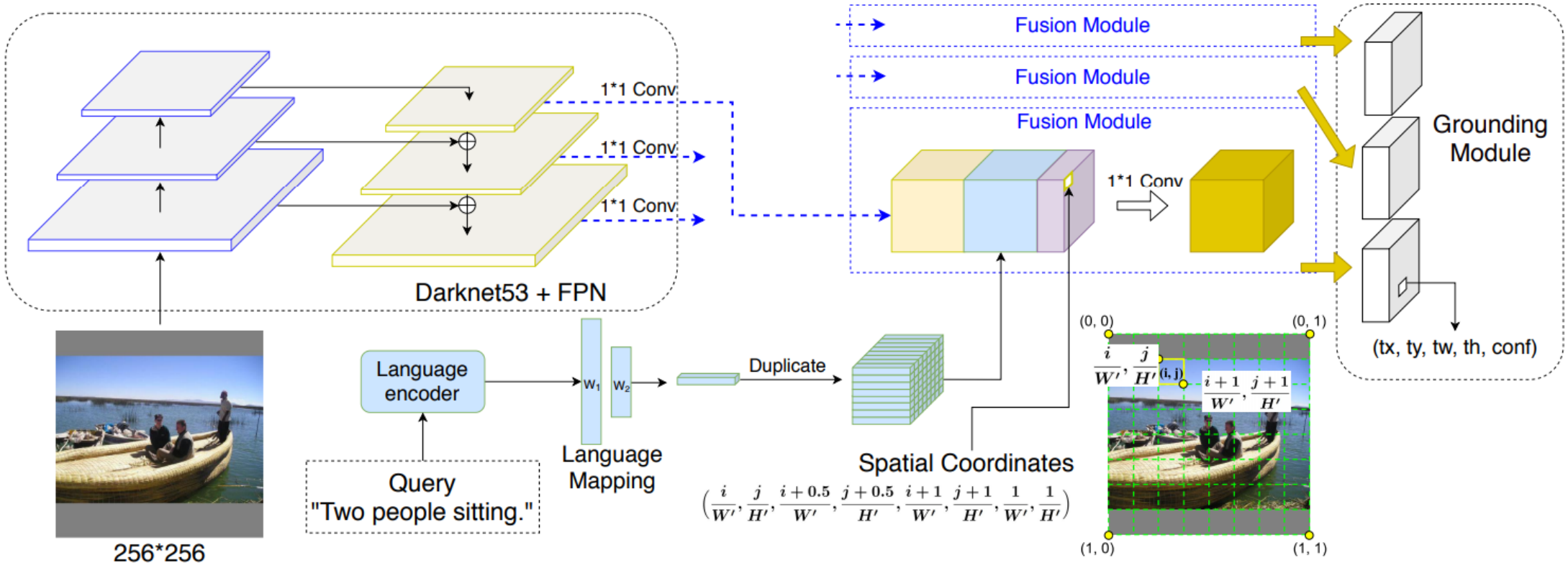
Two-stage methods

- Rely on proposal network
- Computation spent



ICCV 2019, A Fast and Accurate One-Stage Approach to Visual Grounding

One-Stage Method



ICCV 2019, A Fast and Accurate One-Stage Approach to Visual Grounding

One-Stage Method

Table 1. Phrase localization results on the test set of Flickr30K Entities [30].

Method	Region Proposals	Visual Features	Language Embedding	Accu@0.5	Time (ms)
SCRC [14]	Edgebox N=100	VGG16-Imagenet	LSTM	27.80	-
DSPE [43]	Edgebox N=100	VGG19-Pascal	Word2vec, FV	43.89	-
GroundeR [35]	Selec. Search N=100	VGG16-Pascal	LSTM	47.81	-
CCA [30]	Edgebox N=200	VGG19-Pascal	Word2vec, FV	50.89	-
IGOP [44]	None	Multiple Network	N-hot	53.97	-
MCB + Reg + Spatial [2]	Selec. Search N=100	VGG16-Pascal	LSTM	51.01	-
MNN + Reg + Spatial [2]	Selec. Search N=100	VGG16-Pascal	LSTM	55.99	-
Similarity Net [42]	Edgebox N=200	VGG19-Pascal	Word2vec, FV	51.05	-
Similarity Net by CITE [29]	Edgebox N=200	VGG16-Pascal	Word2vec, FV	54.52	-
CITE [29]	Edgebox N=500	VGG16-Pascal	Word2vec, FV	59.27	-
CITE [29]	Edgebox N=500	VGG16-Flickr30K	Word2vec, FV	61.89	-
Similarity Net-Resnet [42]	Edgebox N=200	Res101-COCO	Word2vec, FV	60.89	184
CITE-Resnet [29]	Edgebox N=200	Res101-COCO	Word2vec, FV	61.33	196
Similarity Net-Darknet [42]	Edgebox N=200	Darknet53-COCO	Word2vec, FV	41.04	305
Ours-FV	None	Darknet53-COCO	Word2vec, FV	68.38	**16**
Ours-LSTM	None	Darknet53-COCO	LSTM	67.62	21
Ours-Bert-no Spatial	None	Darknet53-COCO	Bert	67.08	38
Ours-Bert	None	Darknet53-COCO	Bert	**68.69**	38
ECCV20 Sub-Query				69.28	
G3RAPHGROUND++				66.93	

ICCV 2019, A Fast and Accurate One-Stage Approach to Visual Grounding

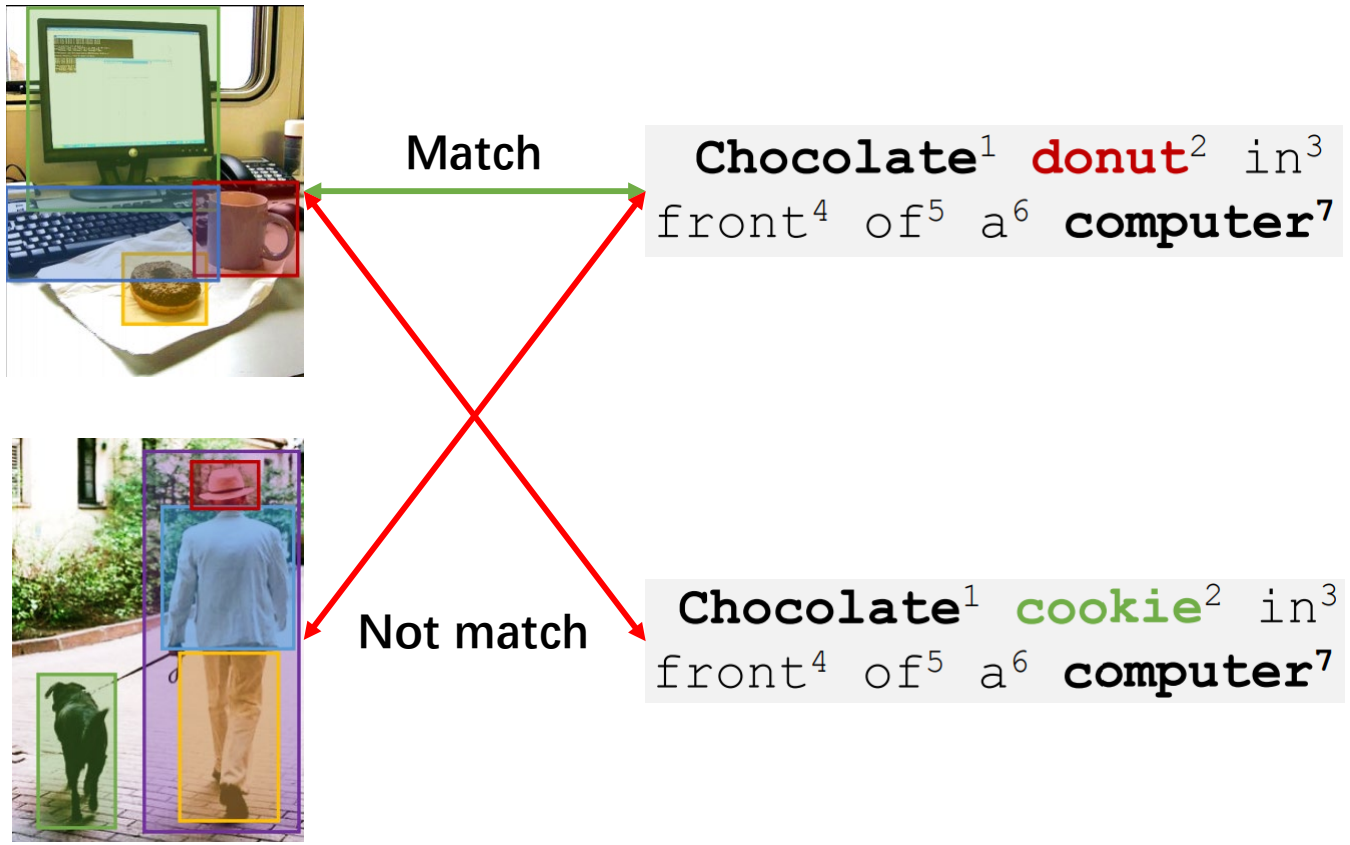
One-Stage Method

Table 2. Referring expression comprehension results on the test set of ReferItGame [15].

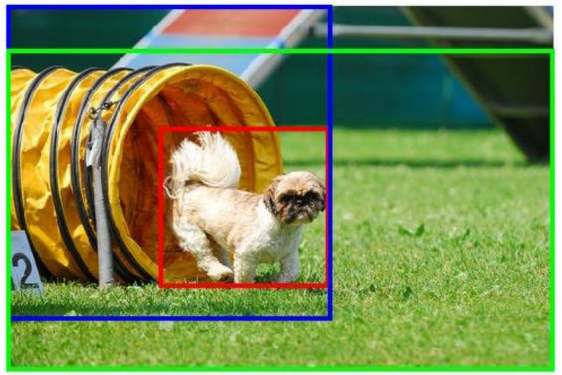
Method	Region Proposals	Visual Features	Language Embedding	Accu@0.5	Time (ms)
SCRC [14]	Edgebox N=100	VGG16-Imagenet	LSTM	17.93	-
GroundeR + Spacial [35]	Edgebox N=100	VGG16-Pascal	LSTM	26.93	-
VC [50]	SSD Detection [21]	VGG16-COCO	LSTM	31.13	-
CGRE [23]	Edgebox	VGG16	LSTM	31.85	-
MCB + Reg + Spatial [2]	Edgebox N=100	VGG16-Pascal	LSTM	26.54	-
MNN + Reg + Spatial [2]	Edgebox N=100	VGG16-Pascal	LSTM	32.21	-
Similarity Net by CITE [29]	Edgebox N=500	VGG16-Pascal	Word2vec, FV	31.26	-
CITE [29]	Edgebox N=500	VGG16-Pascal	Word2vec, FV	34.13	-
IGOP [44]	None	Multiple Network	N-hot	34.70	-
Similarity Net-Resnet [42]	Edgebox N=200	Res101-COCO	Word2vec, FV	34.54	184
CITE-Resnet [29]	Edgebox N=200	Res101-COCO	Word2vec, FV	35.07	196
Similarity Net-Darknet [42]	Edgebox N=200	Darknet53-COCO	Word2vec, FV	22.37	305
Ours-FV	None	Darknet53-COCO	Word2vec, FV	59.18	**16**
Ours-LSTM	None	Darknet53-COCO	LSTM	58.76	21
Ours-Bert-no Spatial	None	Darknet53-COCO	Bert	58.16	38
Ours-Bert	None	Darknet53-COCO	Bert	**59.30**	38
ECCV20 Sub-Query				64.60	
G3RAPHGROUND++				44.91	

Weakly Supervised Visual Grounding

Task Definition: Only Image-Text pairs are provided

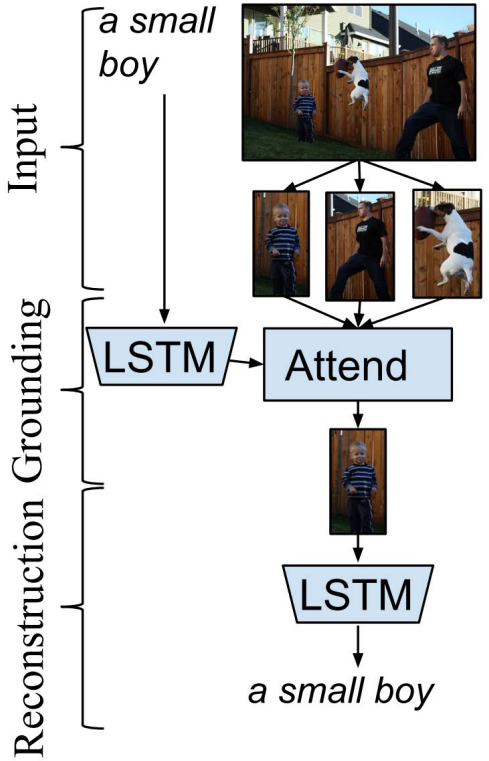


Training with Phrase Reconstruction

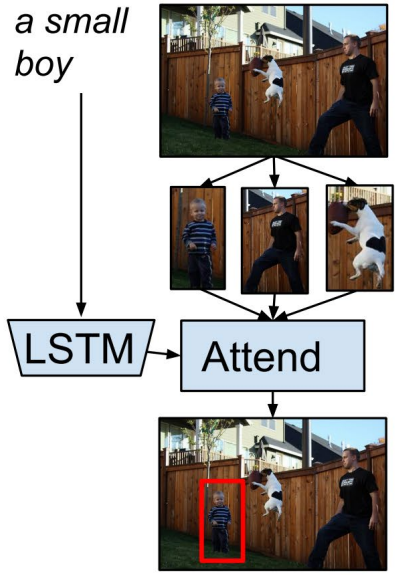


A little brown and white dog emerges from a yellow collapsable toy tunnel onto the lawn.

(a) Predicted grounding.



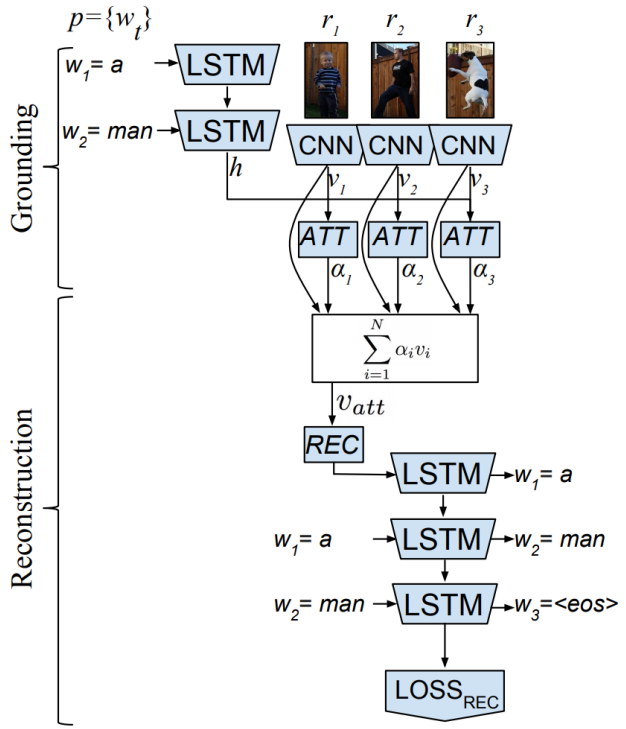
(b) Training time.



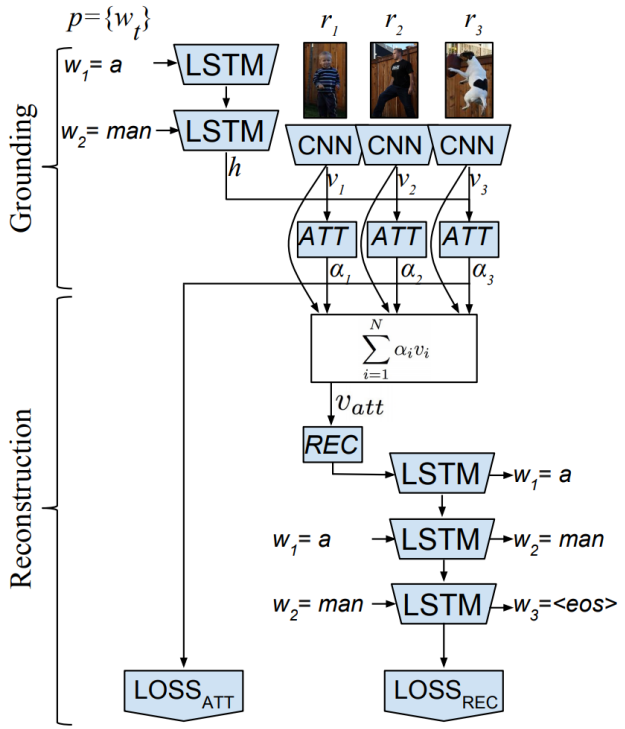
(c) Test time.

Methods

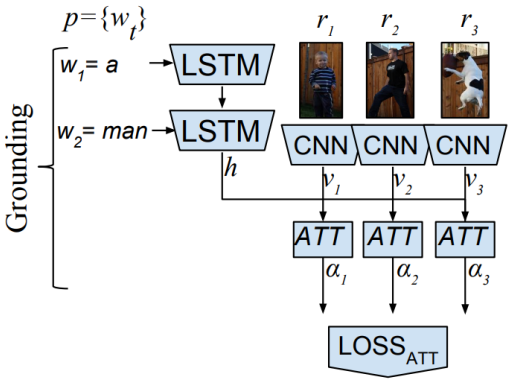
Training with Phrase Reconstruction



(a) Unsupervised



(b) Semi-supervised



(c) Fully supervised

ECCV 2016, Grounding of Textual Phrases in Images by Reconstruction

Training with Phrase Reconstruction

Approach	Accuracy		
	Other	VGG-CLS	VGG-DET
Unsupervised training			
Deep Fragments [6]	21.78	-	-
GroundeR	-	24.66	28.94
Supervised training			
CCA [35]	-	27.42	-
SCRC [18]	-	27.80	-
DSPE [45]	-	-	43.89
GroundeR	-	41.56	47.81
Semi-supervised training			
GroundeR 3.12% annot.	-	33.02	42.32
GroundeR 6.25% annot.	-	37.10	44.02
GroundeR 12.5% annot.	-	38.67	44.96
GroundeR 25.0% annot.	-	39.31	45.32
GroundeR 50.0% annot.	-	40.72	46.65
GroundeR 100.0% annot.	-	42.43	48.38
Proposal upperbound	77.90	77.90	77.90

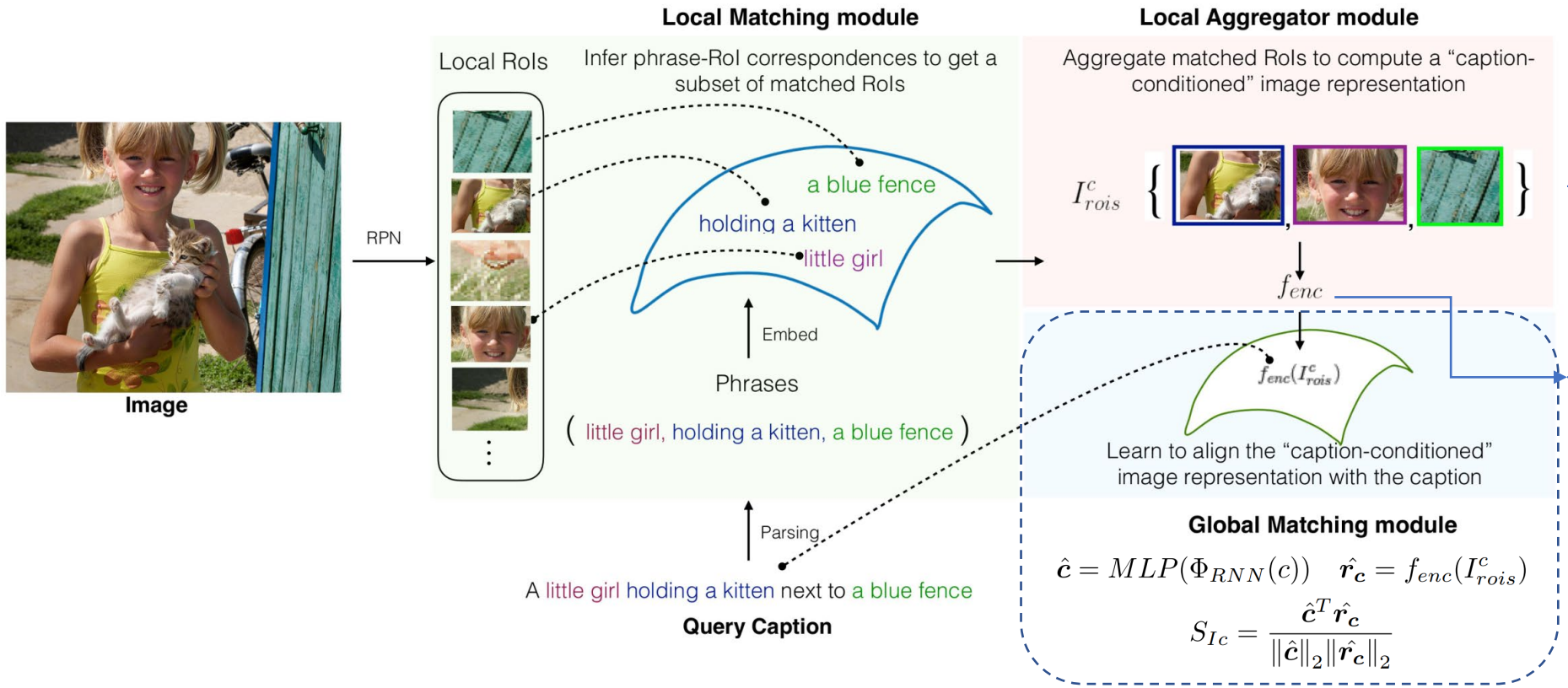
Flickr 30k Entities

Approach	Accuracy		
	Other	VGG	VGG+SPAT
Unsupervised training			
LRCN [9] (reported in [18])	8.59	-	-
CAFFE-7K [15] (reported in [18])	10.38	-	-
GroundeR	-	10.69	10.70
Supervised training			
SCRC [18]	-	-	17.93
GroundeR	-	23.44	26.93
Semi-supervised training			
GroundeR 3.12% annot.	-	13.70	15.03
GroundeR 6.25% annot.	-	16.19	19.53
GroundeR 12.5% annot.	-	19.02	21.65
GroundeR 25.0% annot.	-	21.43	24.55
GroundeR 50.0% annot.	-	22.67	25.51
GroundeR 100.0% annot.	-	24.18	28.51
Proposal upperbound	59.38	59.38	59.38

ReferItGame

Methods

Use Caption-Conditioned Image Feature



Choosing strategy for each phrase:

- argmax
- top-k
- attention

- order-invariant (MLP+mean)
- order-dependent (GRU)

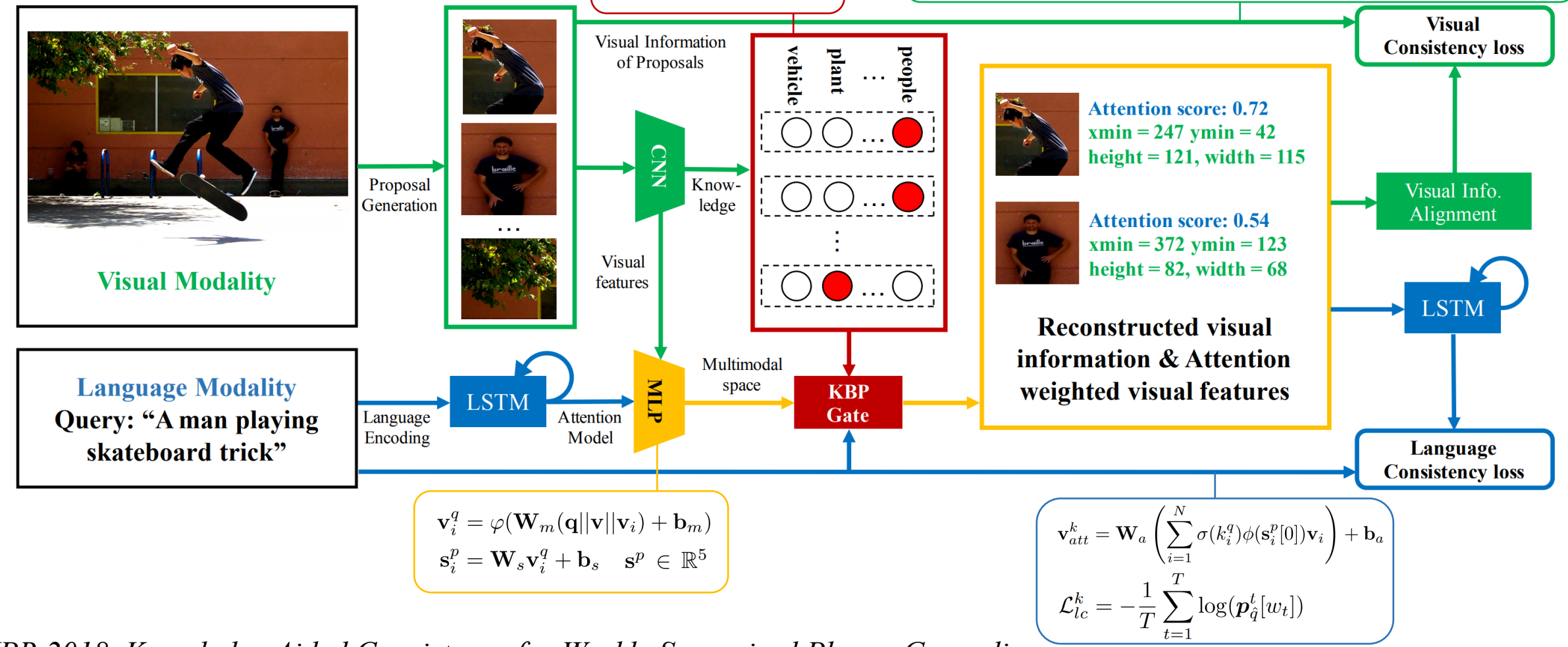
ICCV 2019, Align2ground: Weakly supervised phrase grounding guided by image-caption alignment

Use Caption-Conditioned Image Feature

			COCO					Flickr30k				
			Caption-to-Image retrieval				Phrase	Caption-to-Image retrieval				Phrase
			$R@1$	$R@5$	$R@10$	Med r	$Det.\%$	$R@1$	$R@5$	$R@10$	Med r	$Det.\%$
Global			39.3	74.8	86.3	2	12.2	27.1	56.0	68.4	4	8.0
Pooling-based (words)			47.9	81.7	91.0	2	10.7	40.7	71.2	80.9	2	8.4
Pooling-based (phrases)			48.4	81.7	91.2	2	10.8	41.4	71.4	81.2	2	8.9
Align2Ground												
Proposed model	permInv	max	40.3	76.3	87.8	2	14.5	29.1	60.8	72.7	3	11.5
		topk	56.6	84.9	92.8	1	14.7	49.7	74.8	83.3	2	11.2
		attention	42.8	78.1	89.1	2	10.2	37.9	67.0	77.8	2	6.2
	sequence	max	39.4	75.0	87.1	2	14.5	29.9	60.9	72.7	3	11.5
		topk	58.4	86.1	93.5	1	14.5	47.9	75.6	83.5	2	11.3
		attention	41.9	77.1	88.4	2	9.8	38.2	68.4	78.2	2	5.6

Table 1: Phrase localization and Caption-to-Image retrieval results for models trained on COCO and Flickr30k datasets. Note that we report phrase localization numbers on VisualGenome in all the cases. We compare our proposed model (*permInv-topk*) with two prior methods and with different choices for the Local Matching module (max/topk/attention) and the Local Aggregator module (permInv/sequence) as discussed in Section 3.

Incorporate Reconstruction Task



CVPR 2018, Knowledge Aided Consistency for Weakly Supervised Phrase Grounding

Incorporate Reconstruction Task

Phrase Type	people	clothing	body parts	animals	vehicles	instruments	scene	other
GroundeR (VGG_{det}) [34]	44.32	**9.02**	0.96	46.91	46.00	19.14	28.23	16.98
LC + Soft KBP	55.23	4.21	2.49	67.18	54.50	11.73	37.37	13.25
VC + Soft KBP	51.56	5.33	2.87	58.11	51.50	20.01	26.86	12.63
KAC Net (Hard KBP)	55.14	7.29	2.68	73.94	66.75	20.37	43.14	17.05
KAC Net (Soft KBP)	**58.42**	7.63	**2.97**	**77.80**	**69.00**	**20.37**	**43.53**	**17.05**

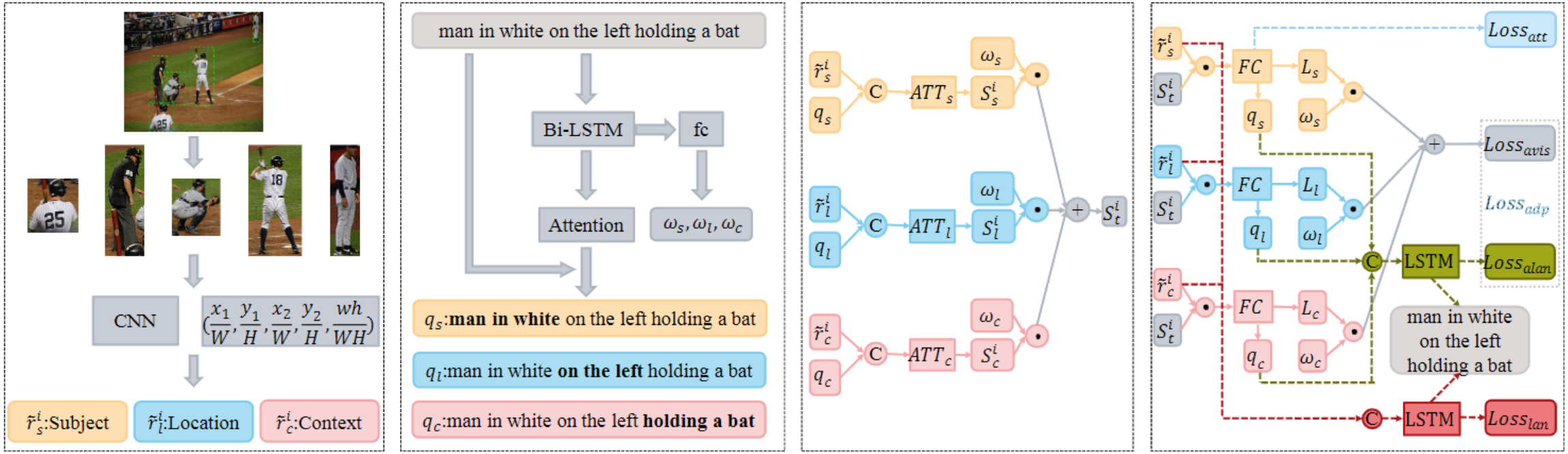
Table 3. Phrase grounding performances for different phrase types defined in Flickr30K Entities. Accuracy is in percentage.

Approach	Accuracy (%)
Compared approaches	
GroundeR (LC) (VGG_{cls}) [34]	24.66
GroundeR (LC) (VGG_{det}) [34]	28.93
Our approaches	
VC + Hard KBP (VGG_{det})	28.58
VC + Soft KBP (VGG_{det})	30.60
LC + Hard KBP (VGG_{det})	32.17
LC + Soft KBP (VGG_{det})	34.31
KAC Net + Hard KBP (VGG_{det})	37.41
KAC Net + Soft KBP (VGG_{det})	**38.71**

Table 1. Different models' performance on Flickr30K Entities. We explicitly evaluate performance of visual consistency (VC), language consistency (LC) branches with Hard and Soft KBP Gates. We leverage knowledge from MSCOCO [25] classification task.

Methods

Incorporate Reconstruction Task



(a) Visual feature encoding. (b) Language feature encoding. (c) Adaptive grounding. (d) Collaborative Reconstruction.

- Regions' C3/C4 feature
- Regions' absolution/relative position
- C4/relative position feature from surrounding regions with the maximum response to the query

$$L_x = \text{MSE}(v_x, q_x), \quad x \in (s, l, c)$$
$$Loss_{avis} = w_s L_s + w_l L_l + w_c L_c$$

$$Loss_{att} = f_{\text{BCE}}(y_{ij}, p_{ij})$$

$$f_{alan} = \phi_{\text{ReLU}}(W_l([q_s, q_l, q_c]) + b_l)$$
$$Loss_{alan} = -\frac{1}{B}\sum_{b=1}^{B}\log(P(\hat{q}|f_{alan}))$$

$$f_{vis} = \sum_{i=1}^{N} S_t^i r_{vis}^i$$
$$Loss_{lan} = -\frac{1}{B}\sum_{b=1}^{B}\log(P(\hat{q}|f_{vis}))$$

ICCV 2019, Adaptive reconstruction network for weakly supervised referring expression grounding

Incorporate Reconstruction Task

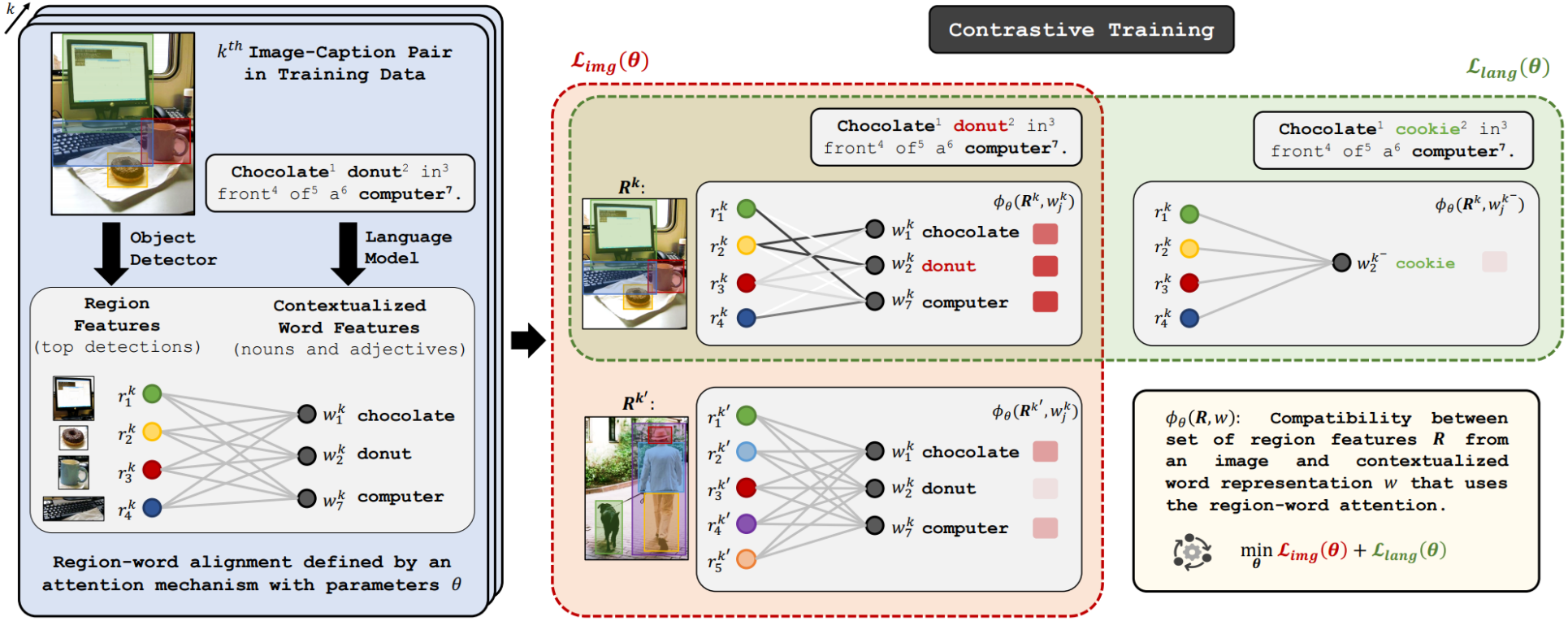
Table 1. Accuracy (IoU $>$ 0.5) on RefCOCO dataset. **Bond**: best result. Red: second best result. Blue: best result of VC.

Methods	Settings	RefCOCO			RefCOCO+			RefCOCOg
		val	testA	testB	val	testA	testB	val
VC	w/o reg	-	13.59	21.65	-	18.79	24.14	25.14
VC	-	-	17.34	20.98	-	23.24	24.91	33.79
VC	w/o α	-	33.29	30.13	-	34.60	31.58	30.26
VC (det)	w/o reg	-	17.14	22.30	-	19.74	24.05	28.14
VC (det)	-	-	20.91	21.77	-	25.79	25.54	33.66
VC (det)	w/o α	-	32.68	27.22	-	34.68	28.10	29.65
ARN	$L_{adp} + L_{att}$	33.07	**36.43**	29.09	33.53	**36.40**	29.23	33.19
ARN	$L_{lan} + L_{att}$	**38.05**	35.27	**36.47**	34.51	34.40	**36.12**	**39.62**
ARN	$L_{lan} + L_{adp}$	33.60	35.65	31.48	34.40	35.54	32.60	34.50
ARN (det)	$L_{lan} + L_{adp}$	31.58	35.50	28.32	31.73	34.23	29.35	32.60
ARN	$L_{lan} + L_{adp} + L_{att}$	34.26	36.01	33.07	**34.53**	36.01	33.75	34.66
ARN (det)	$L_{lan} + L_{adp} + L_{att}$	32.17	35.35	30.28	32.78	34.35	32.13	33.09

ICCV 2019, Adaptive reconstruction network for weakly supervised referring expression grounding

Methods

Contrastive Learning



InfoNCE Loss

$$\mathcal{L}_k(\theta) = \mathbb{E}_{\mathcal{B}}\left[-\log\left(\frac{e^{\phi_\theta(x,y)}}{e^{\phi_\theta(x,y)} + \sum_{i=1}^{k-1} e^{\phi_\theta(x_i',y)}}\right)\right]$$

BERT negative example generation

$$\underset{s'}{\text{argmax}} \frac{p(s'|c)}{q(s'|s,c)}$$

Caption with word s masked

Original caption

ECCV 2020, Contrastive Learning for Weakly Supervised Phrase Grounding

Contrastive Learning

Flickr30K Entities

Method	Training Data	Visual Features	R@1	R@5	R@10	Accuracy
GroundeR (2015) [33]	Flickr30K Entities	VGG-det (VOC)	28.94	-	-	-
Yeh *et al.*(2018) [44]	Flickr30K Entities	VGG-cls (IN)	22.31	-	-	-
Yeh *et al.*(2018) [44]	Flickr30K Entities	VGG-det (VOC)	35.90	-	-	-
Yeh *et al.*(2018) [44]	Flickr30K Entities	YOLO (COCO)	36.93	-	-	-
KAC Net+Soft KBP (2018) [7]	Flickr30K Entities	VGG-det (VOC)	38.71	-	-	-
Fang *et al.*(2015) [13]	COCO	VGG-cls (IN)	-	-	-	29.00
Akbari *et al.*(2019) [1]	COCO	VGG-cls (IN)	-	-	-	61.66
Akbari *et al.*(2019) [1]	COCO	PNAS Net (IN)	-	-	-	69.19
Align2Ground (2019) [11]	COCO	Faster-RCNN (VG)	-	-	-	71.00
Ours	Flickr30K Entities	Faster-RCNN (VG)	**47.88**	**76.63**	**82.91**	**74.94**
Ours	COCO	Faster-RCNN (VG)	**51.67**	**77.69**	**83.25**	**76.74**

Further Work

- Contrastive Learning
- Grounding through Dialogs
- Visual Query Detection

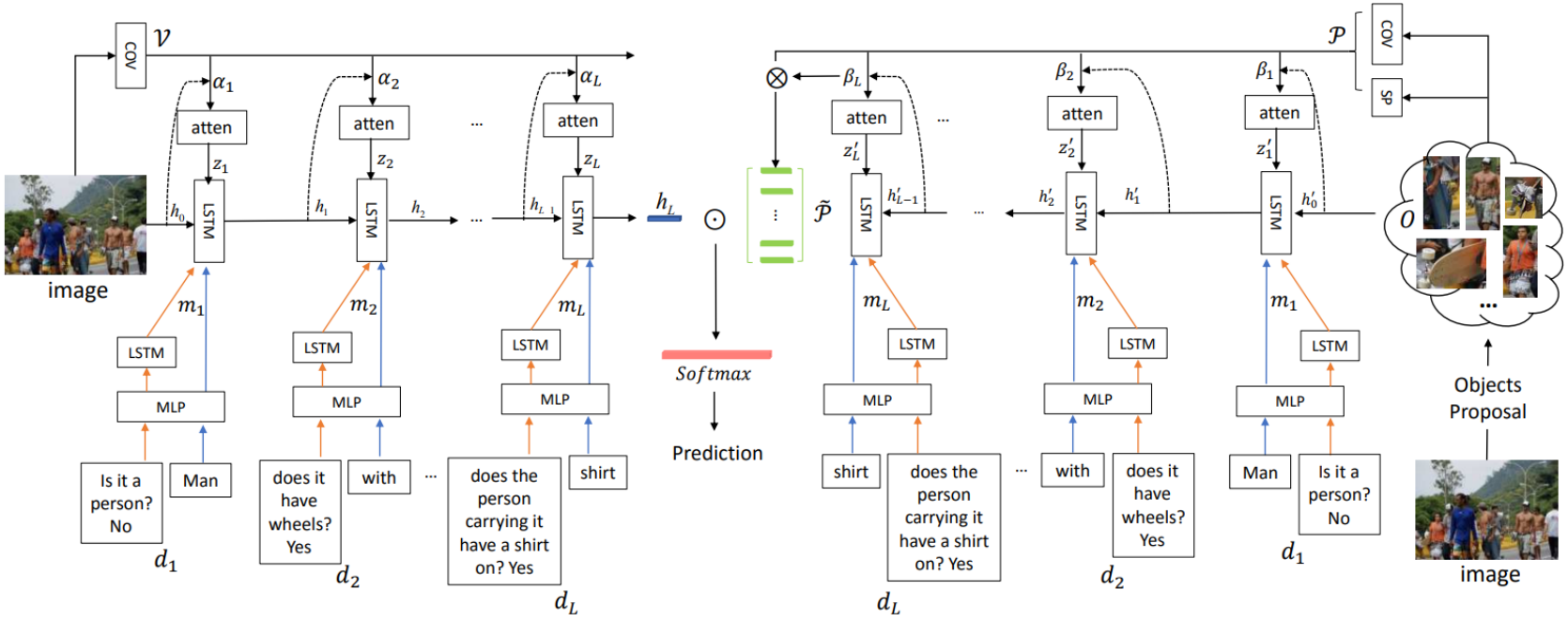
Further Work

Grounding through Dialogs

GuessWhat Dataset

Questioner	Oracle
Is it a vase?	Yes
Is it partially visible?	No
Is it in the left corner?	No
Is it the turquoise and purple one?	Yes

Figure 1: An example game. After a sequence of four questions, it becomes possible to locate the object (highlighted by a green bounding box).

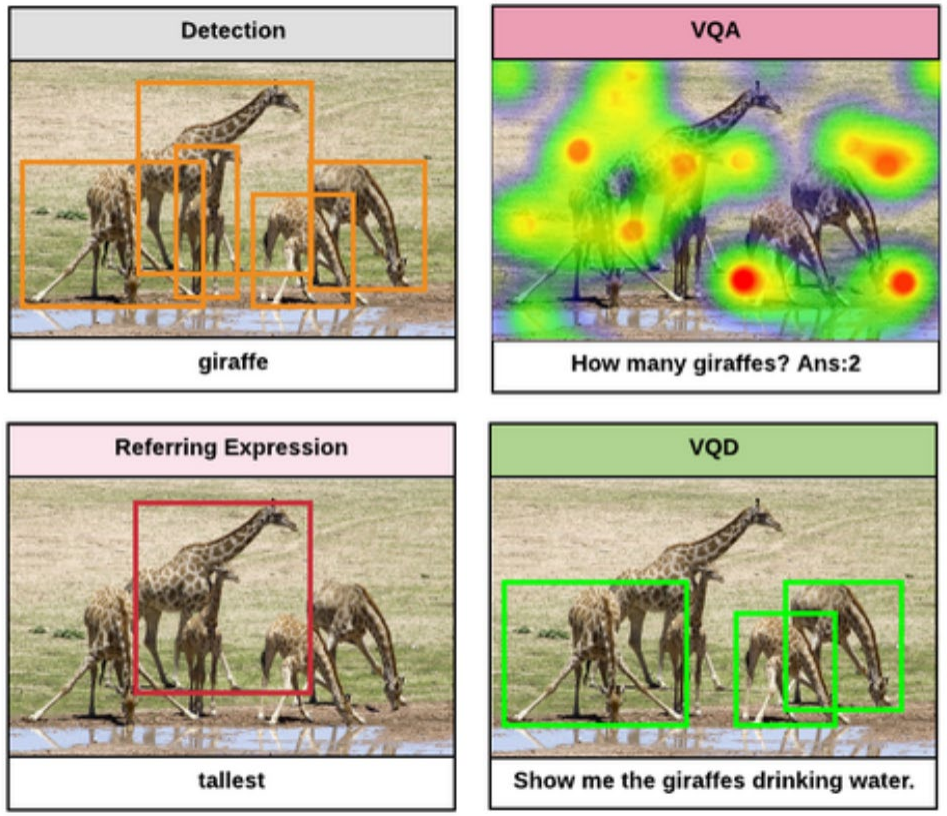


CVPR 2017, GuessWhat?! Visual object discovery through multi-modal dialogue
CVPR 2018, Parallel Attention: A Unified Framework for Visual Object Discovery Through Dialogs and Queries

Further Work

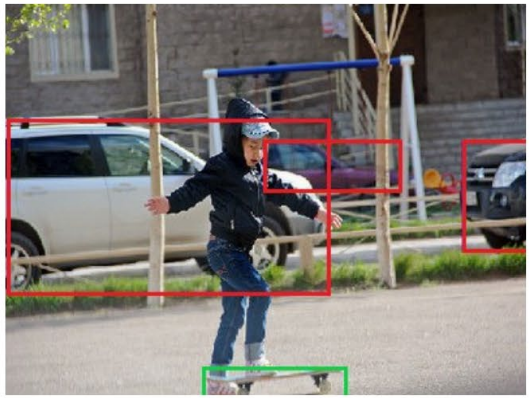
VQD: Visual Query Detection in Natural Scenes

- Object Presence
- Color Reasoning
- Positional Reasoning

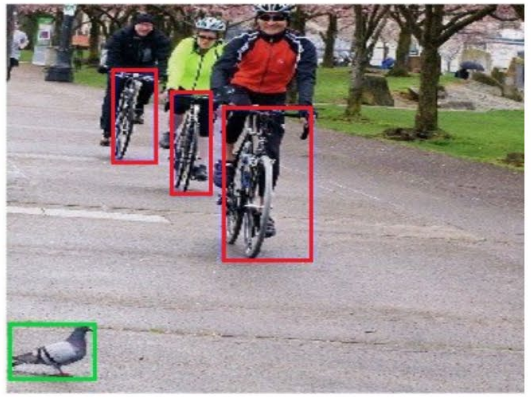


Further Work

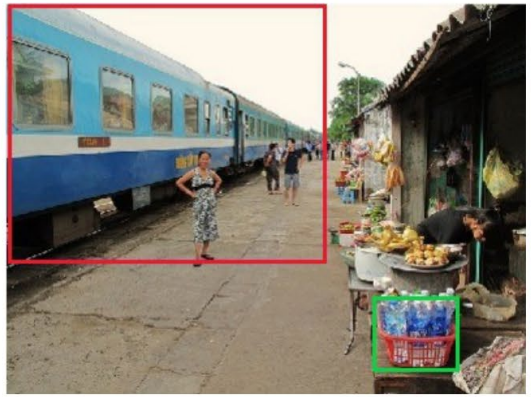
VQD: Visual Query Detection in Natural Scenes



(a) Show the car in the picture. (red) Where is the skateboard in the image? (green)

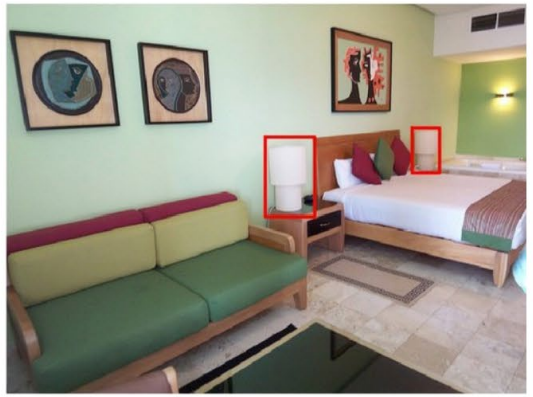


(b) Show me the bicycle. (red) Where is the bird? (green)



(c) Which train is blue in color? (red) Show me the red basket. (green)

(d) Which bin is under the sink? (red) Which bottle is on top of shelf? (green)



(e) Show the lamp beside bed in the image. (red)

(f) Where is the sink in the picture? Where is the toaster in the image?