UNIVERSITY of ROCHESTER

Tencent AI Lab

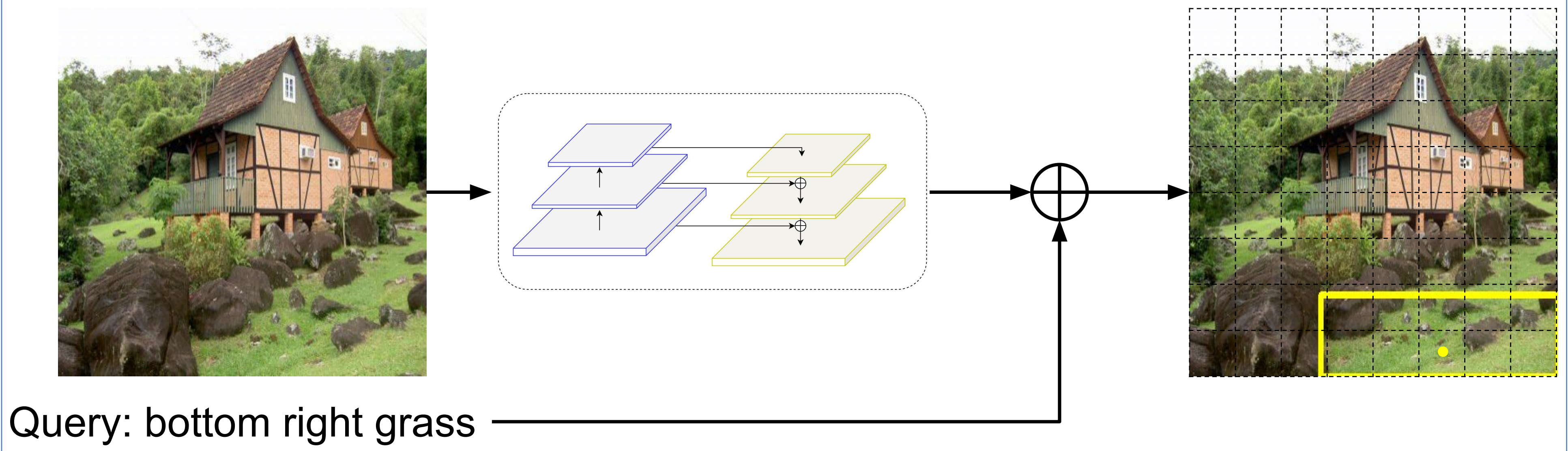
A Fast and Accurate One-Stage Approach to Visual Grounding

Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, Jiebo Luo

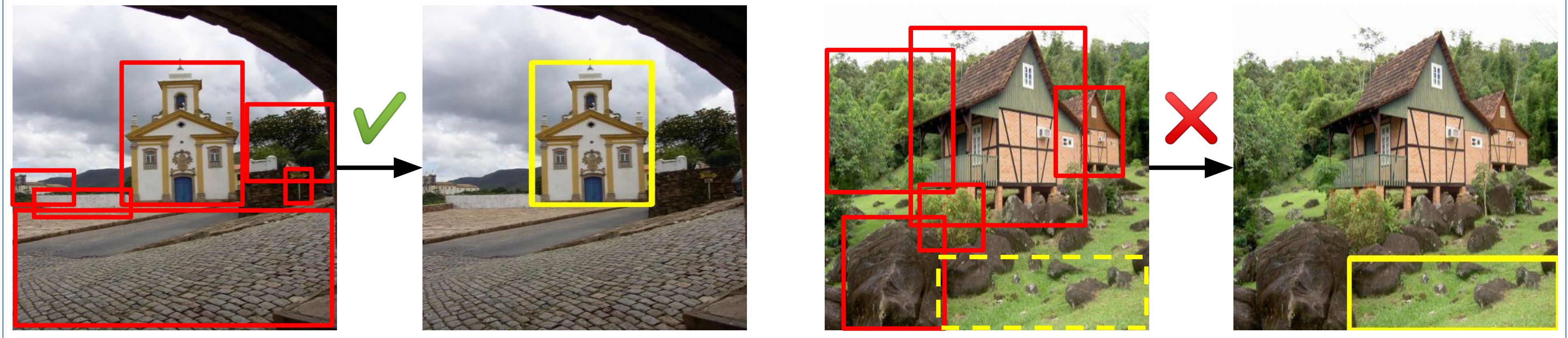
ICCV 2019 Seoul, Korea

Visual Grounding

Ground a natural language query (phrase or sentence) about an image onto a correct region of the image.



Limitations of Propose-and-Rank Methods



Query: center building

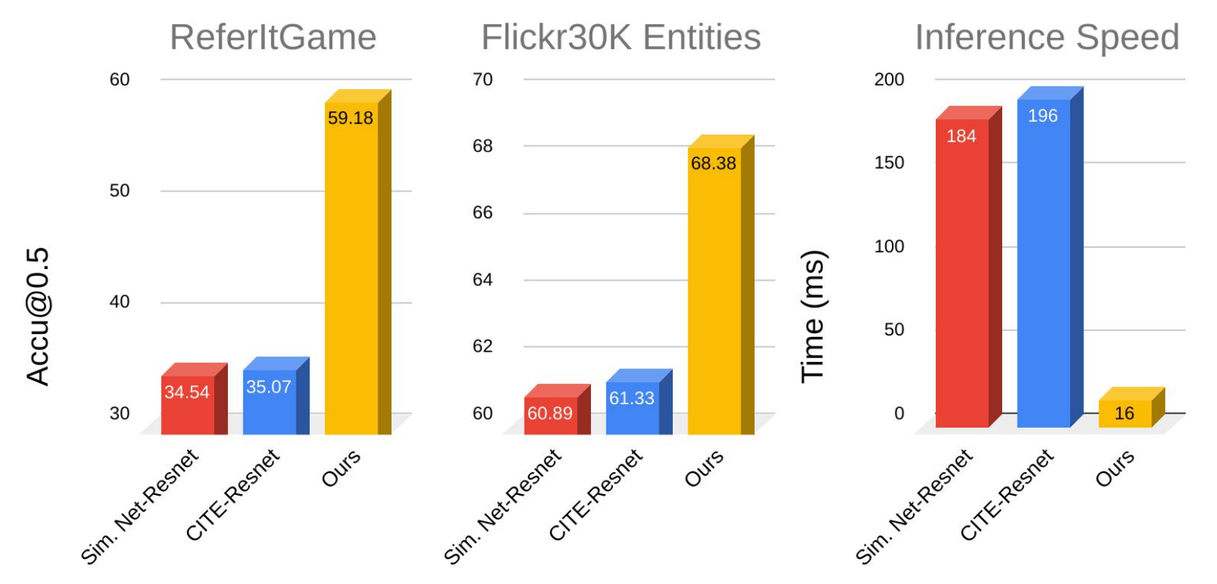
Query: bottom right grass

- Performance: consider limited numbers of candidates
- Speed: slow in getting all candidate features

Contributions

A fast and accurate one-stage approach to visual grounding

- Fast: feature extraction in one pass
- Accurate: consider all possible locations



One-Stage Visual Grounding

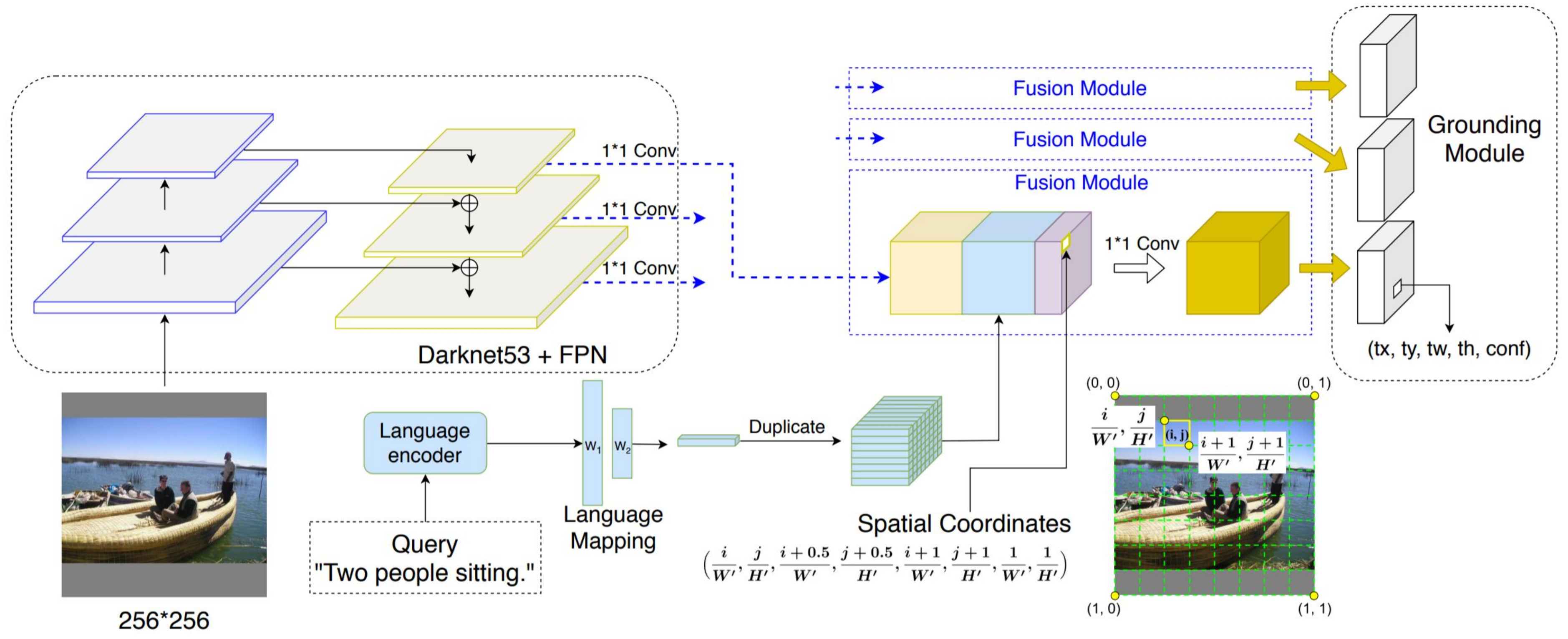


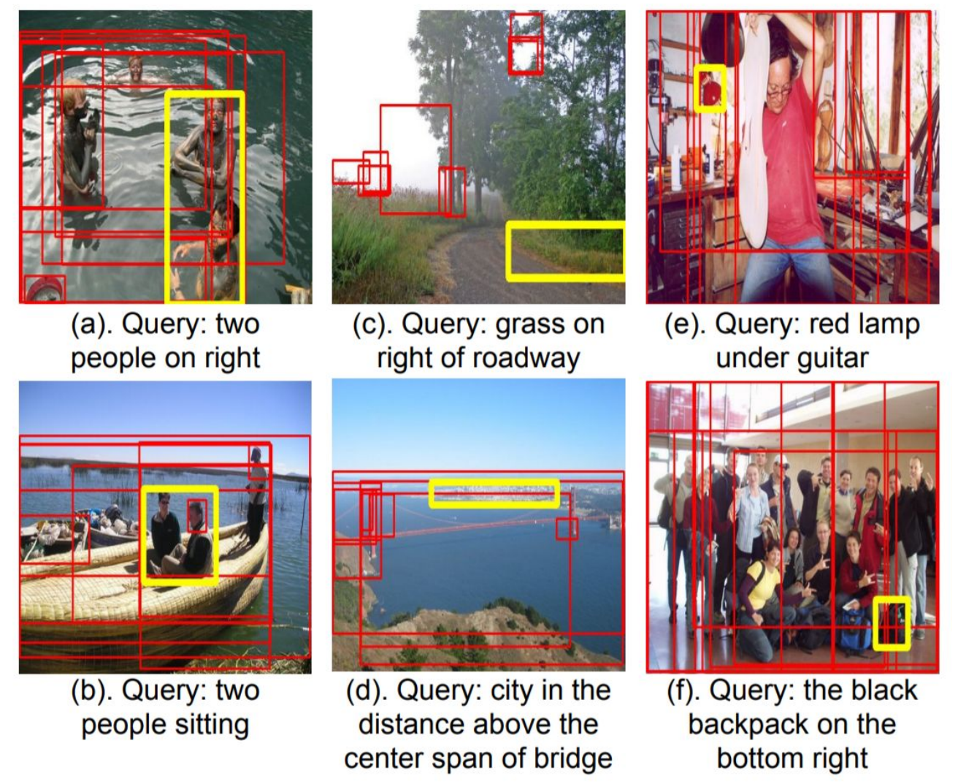
Image-level visual-textual fusion and grounding

- Visual: DarkNet53+FPN
- Textual: Bert, LSTM, Word2vec + Fisher Vector (FV) encoding
- Spatial: Spatial coordinates of the grid

- Fusion Module
  - Concatenation
  - Convolutional layers with 1x1, 3x3 kernels
  - Optionally, cross-location, cross-head attention further improves the performance

- Grounding Module
  - With 256^2 input resolution, $(8^2+16^2+32^2)$ locations * 3 anchors = 4032 predictions
  - Each predicted box consists of the location, size offsets and confidence prediction

Oracle Analyses



(a). Query: two people on right

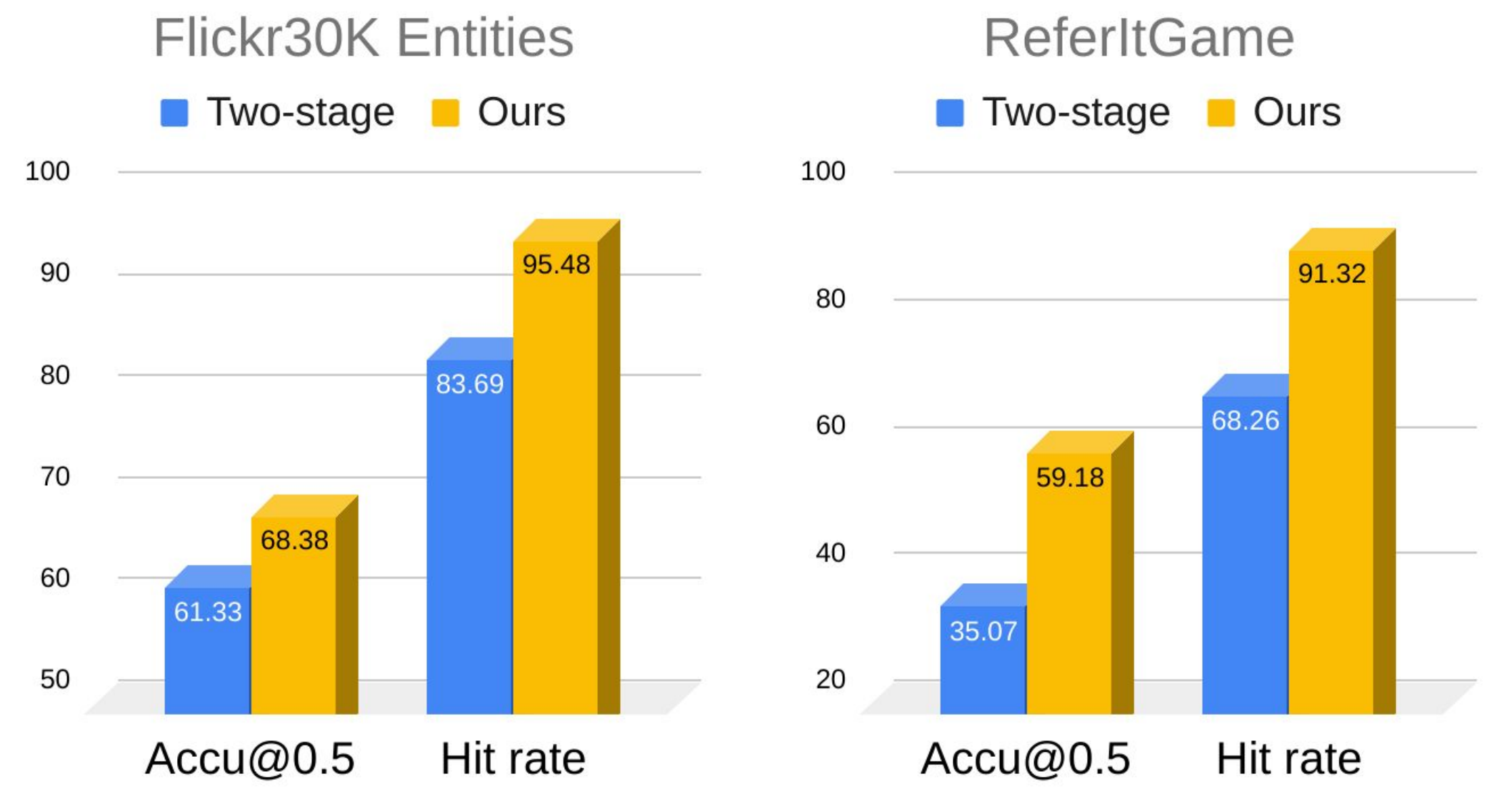
(c). Query: grass on right of roadway

(e). Query: red lamp under guitar

(b). Query: two people sitting

(d). Query: city in the distance above the center span of bridge

(f). Query: the black backpack on the bottom right



Quantitative Results

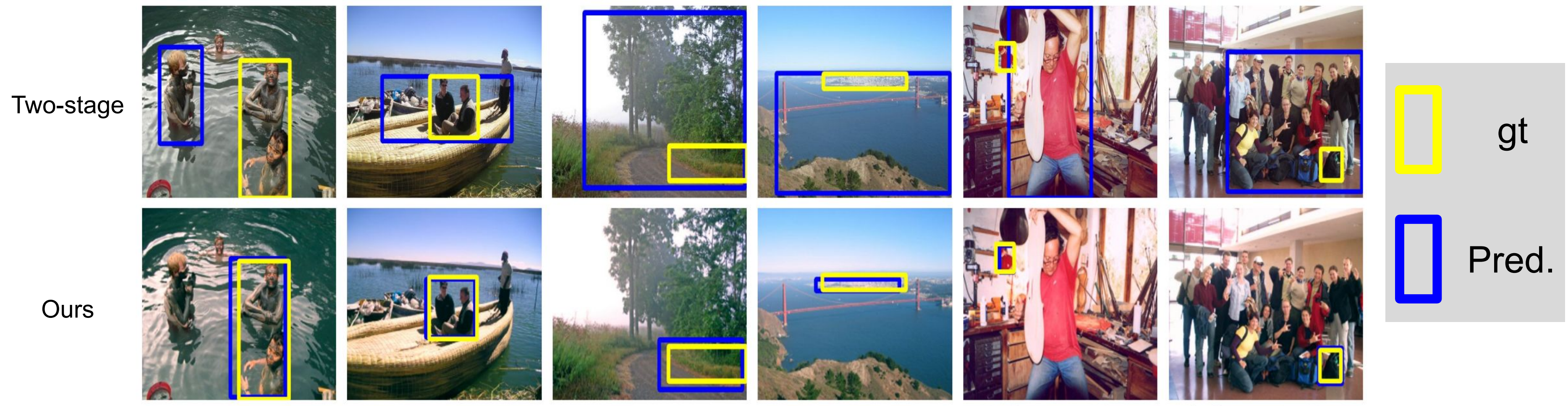
Results on ReferItGame

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

Results on Flickr30K Entites

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

Quantitative Results



(a). Query: two people on right

(b). Query: two people sitting

(c). Query: grass on right of roadway

(d). Query: city in the distance above the center span of bridge

(e). Query: red lamp under guitar

(f). Query: the black backpack on the bottom right