UNIVERSITY of ROCHESTER

Tencent AI Lab

ICCV 2019 Seoul, Korea

A Fast and Accurate One-Stage Approach to Visual Grounding

Zhengyuan Yang

Boqing Gong

Liwei Wang

Wenbing Huang

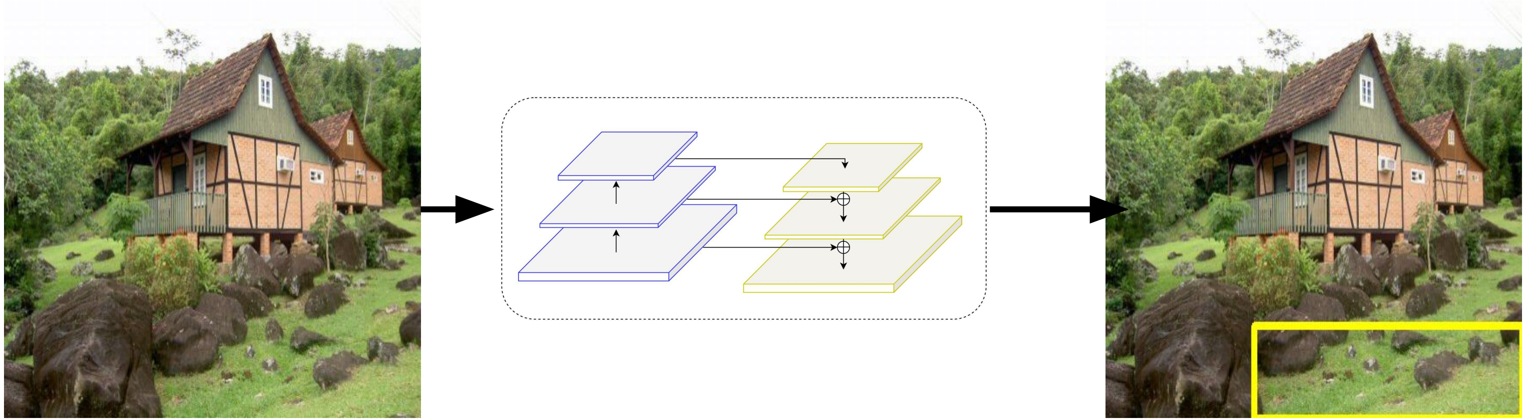
Dong Yu

Jiebo Luo

Presenter: Tianlang Chen

Visual grounding

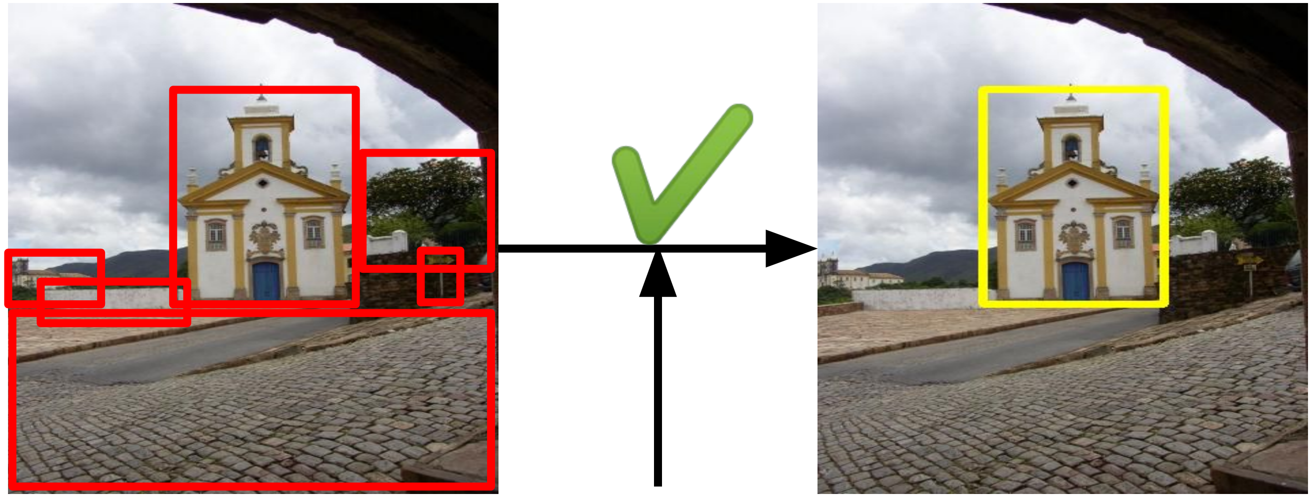
- Grounding a language query onto a region of the image



Query: bottom right grass

Existing framework

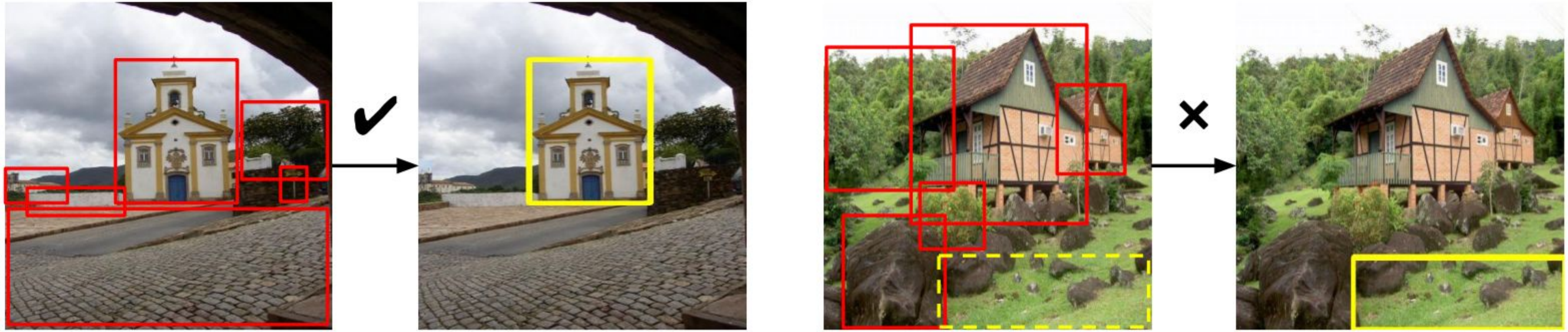
- Two-stage framework



Query: center building

Existing framework

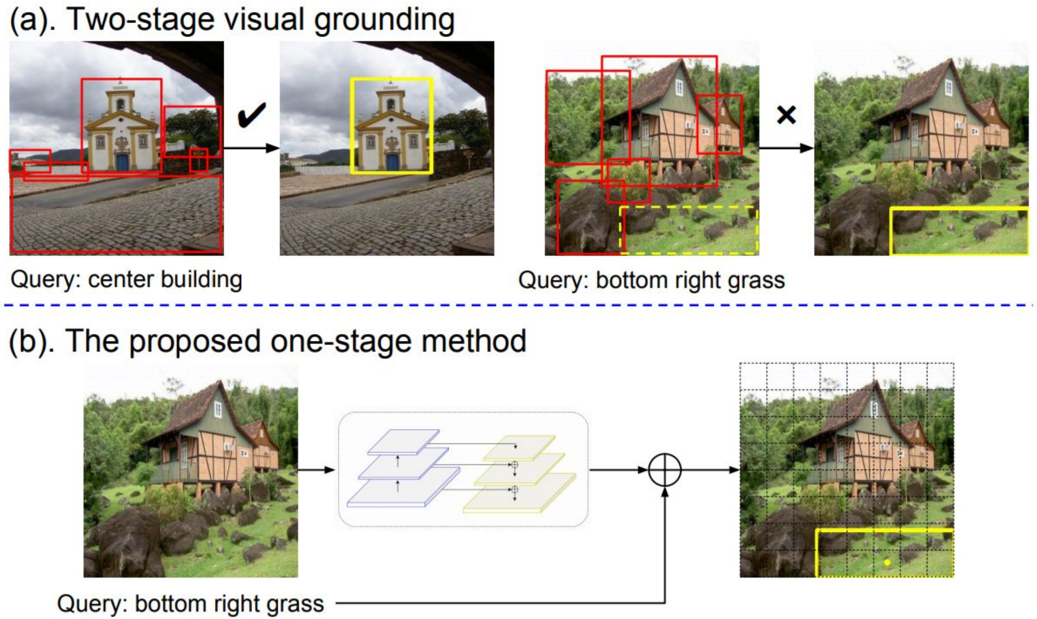
- Performance is capped by the region candidates
- Slow in speed



Query: center building

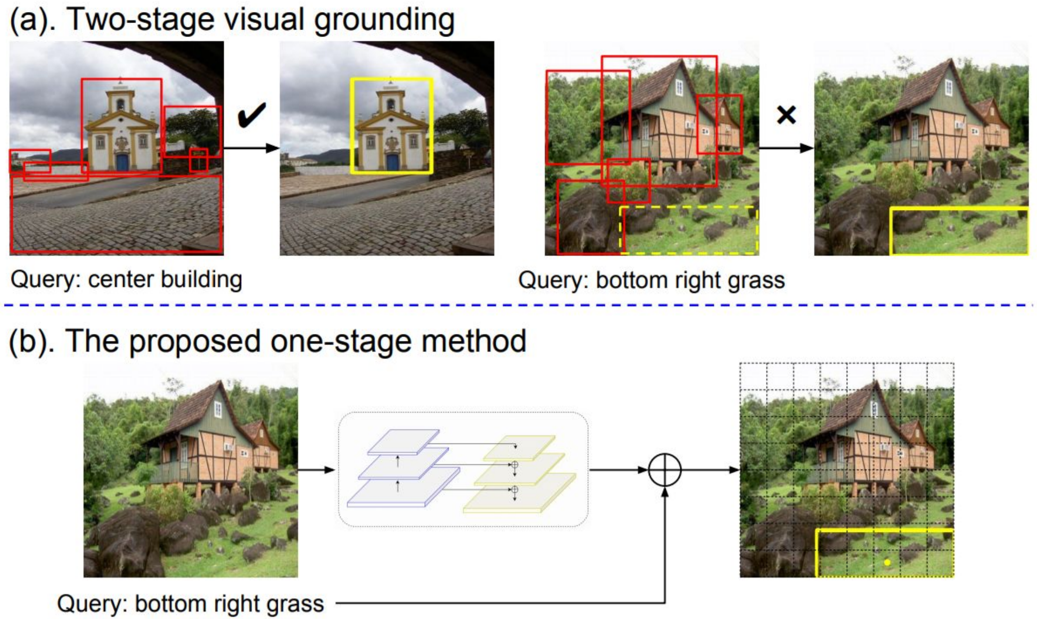
Query: bottom right grass

One-stage visual grounding


(a). Two-stage visual grounding
✔
×
Query: center building
Query: bottom right grass
(b). The proposed one-stage method
Query: bottom right grass

- One-stage approach
- Generally applicable for sub-tasks in grounding

Why one-stage visual grounding

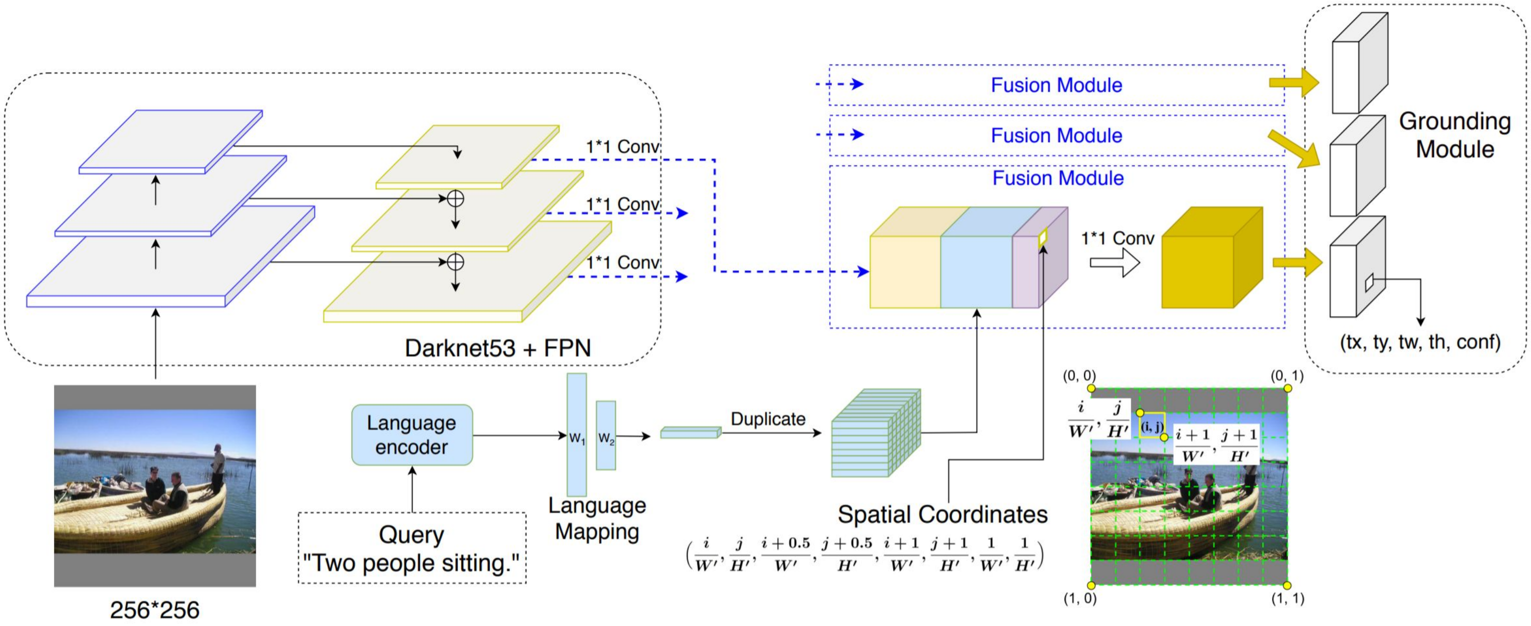

(a). Two-stage visual grounding

(b). The proposed one-stage method

☺

- No region candidates -> 7~20% higher in accuracy
- One-stage -> 10x faster

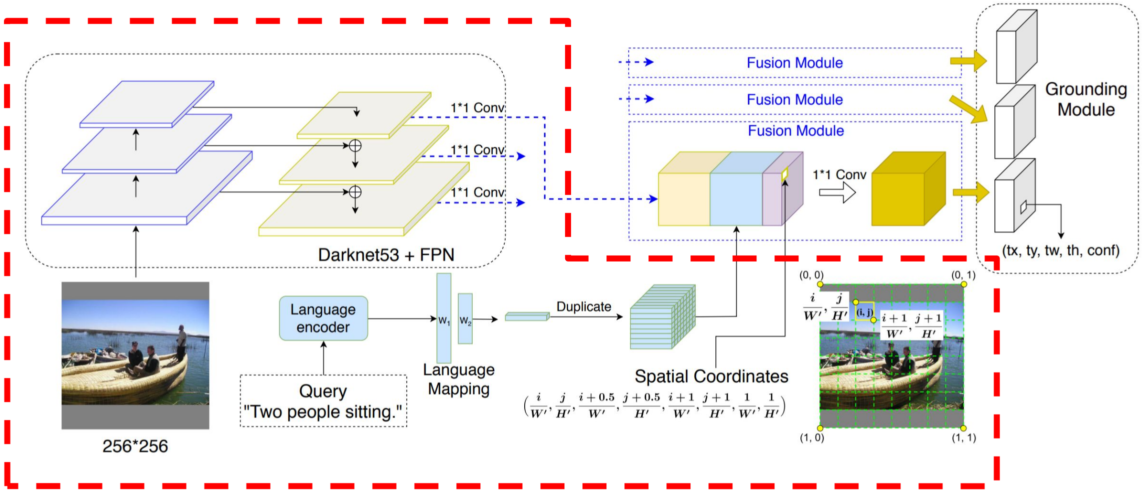
Architecture overview



- Encoder
- Fusion module
- Grounding module

Architecture

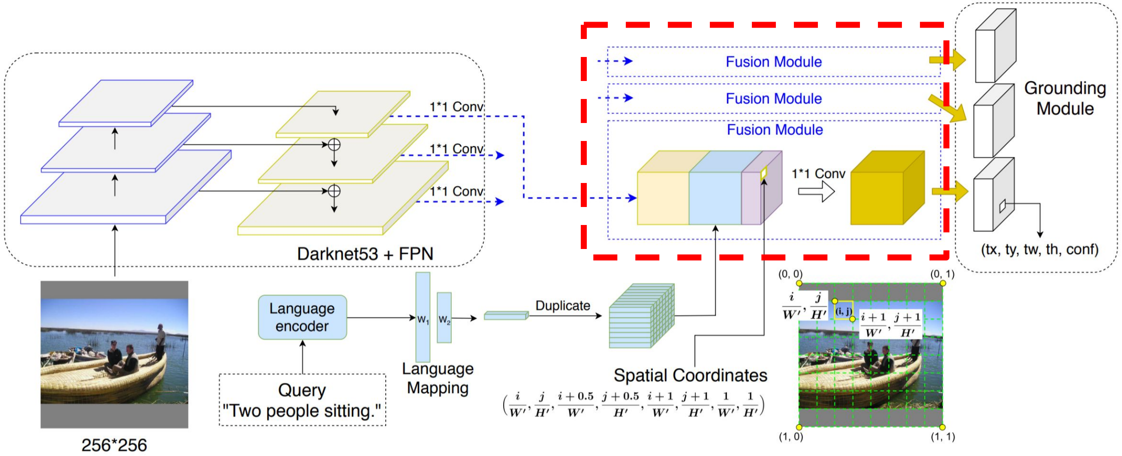
- Encoder
- Fusion module
- Grounding module



- Visual encoder: DarkNet53+FPN
- Language encoder: Bert, LSTM, FV
- Spatial encoder: location related queries

Architecture

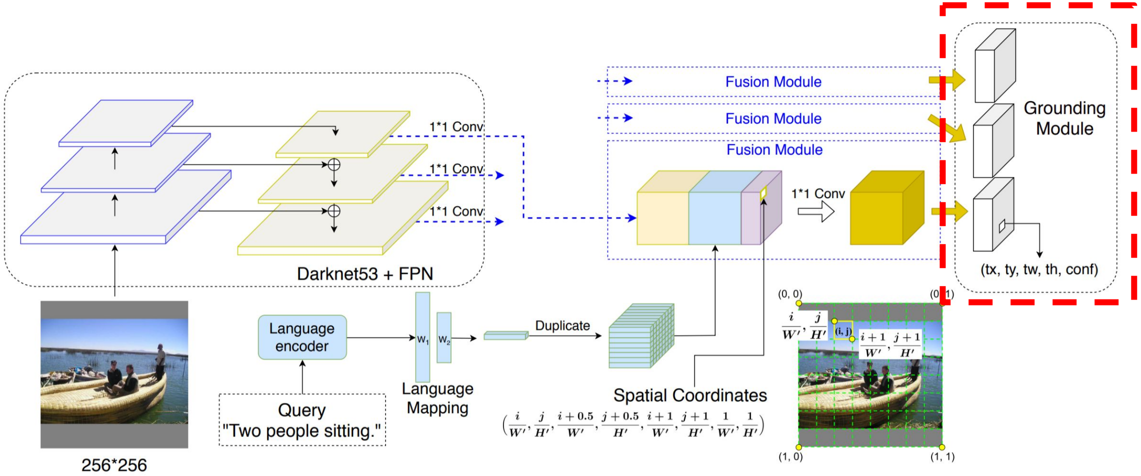
- Encoder
- Fusion module
- Grounding module



- Image-level fusion

Architecture

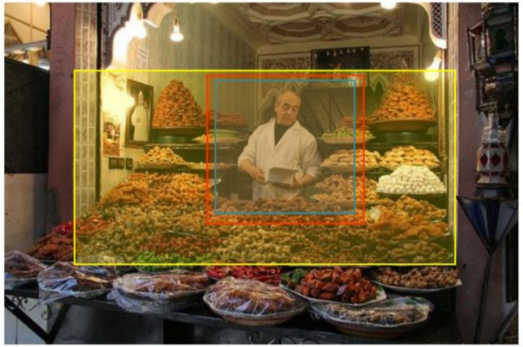
- Encoder
- Fusion module
- Grounding module



- Output format: box + confidence

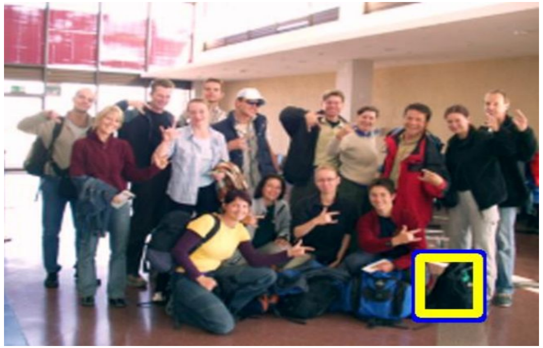
Datasets

- Phrase localization: Flickr 30K Entities
- Referring expression comprehension: ReferItGame



An older man in a white jacket works at a stand featuring a wide variety of colorful food.

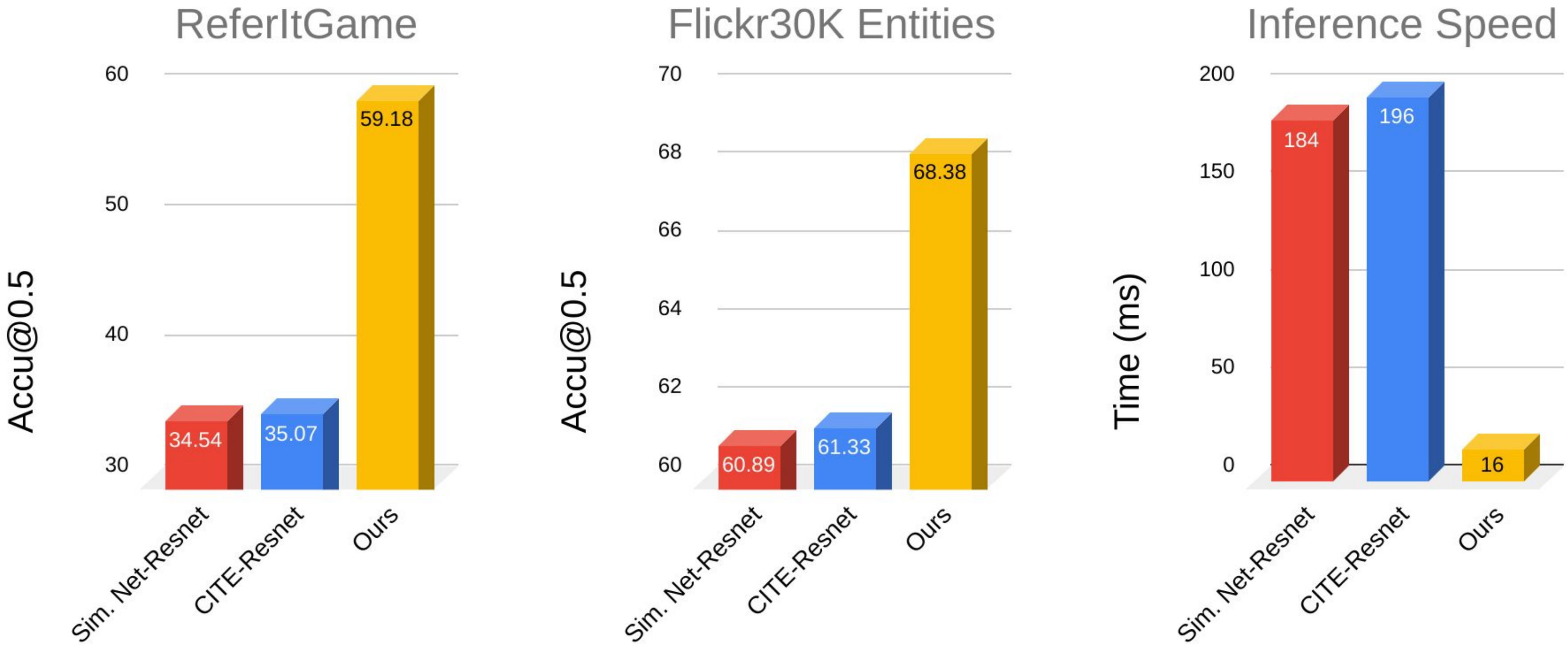
Flickr 30K Entities



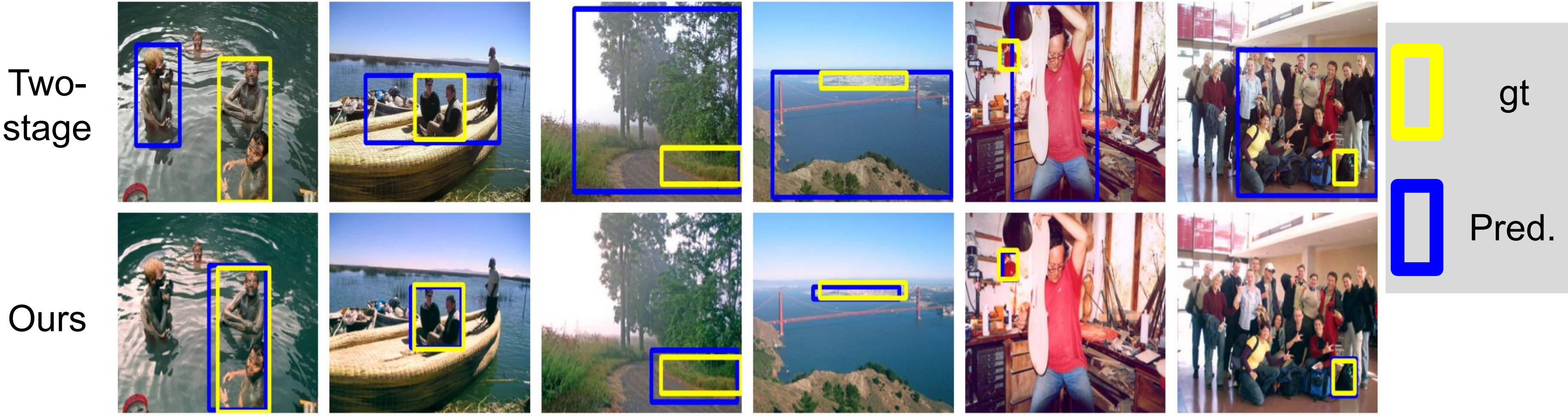
the black backpack on the bottom right

ReferItGame

Comparison to other methods



Qualitative 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

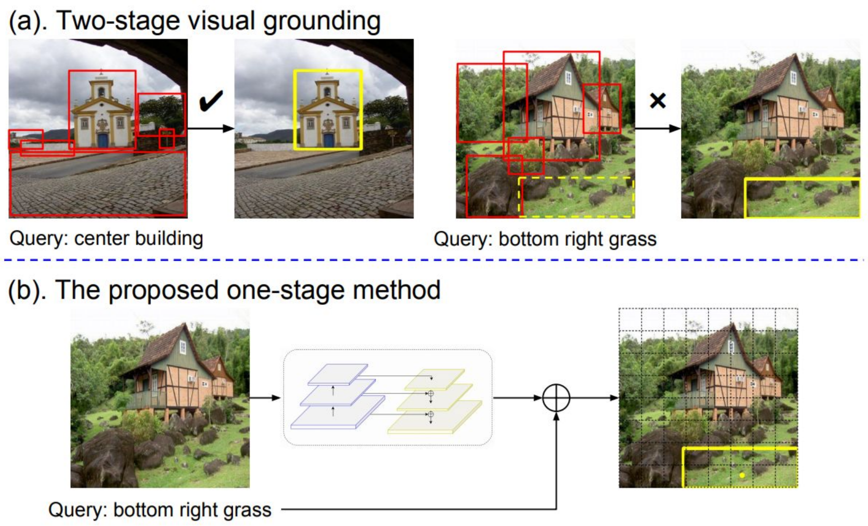
- Union of multiple objects
- Stuff as opposed to things
- Challenging regions

A Fast and Accurate One-Stage Approach to Visual Grounding

Code & models:
https://github.com/zyang-ur/onestage_grounding

Poster: #26

Contact:
zyang39@cs.rochester.edu


(a). Two-stage visual grounding
✔
Query: center building
×
Query: bottom right grass
(b). The proposed one-stage method
Query: bottom right grass