Grounding of Textual Phrases in Images by Reconstruction

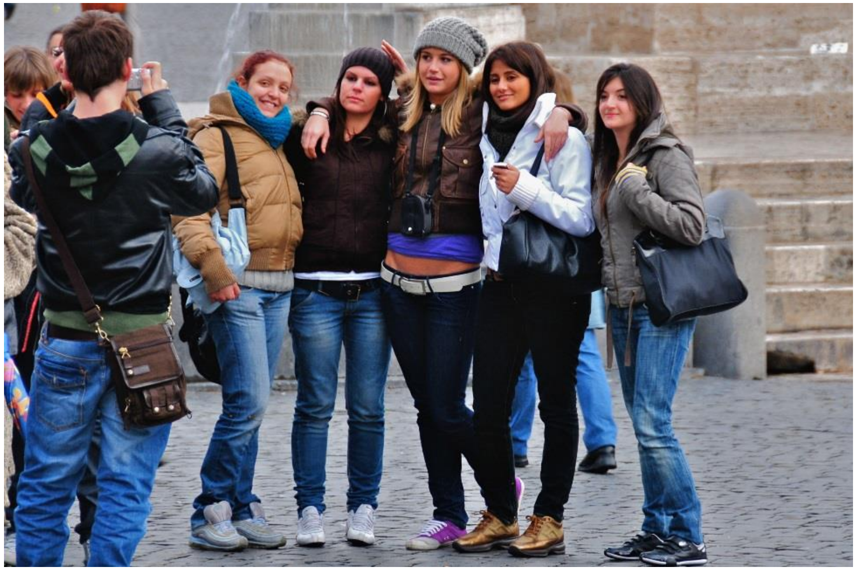
Anna Rohrbach[1]

joint work with
Marcus Rohrbach[2], Ronghang Hu[2],
Trevor Darrell[2], Bernt Schiele[1]

[1] Max Planck Institute for Informatics, Saarland Informatics Campus, Germany
[2] UC Berkeley EECS, Berkeley, CA, United States

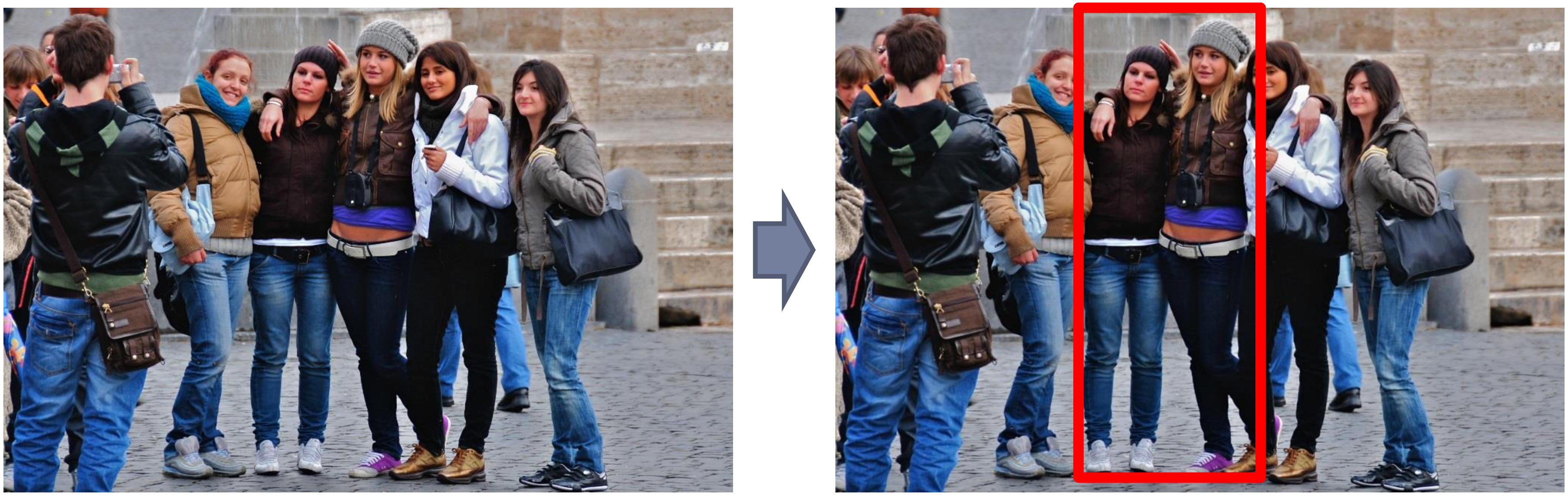
Visual Grounding: Task

The two girls in hats in the middle



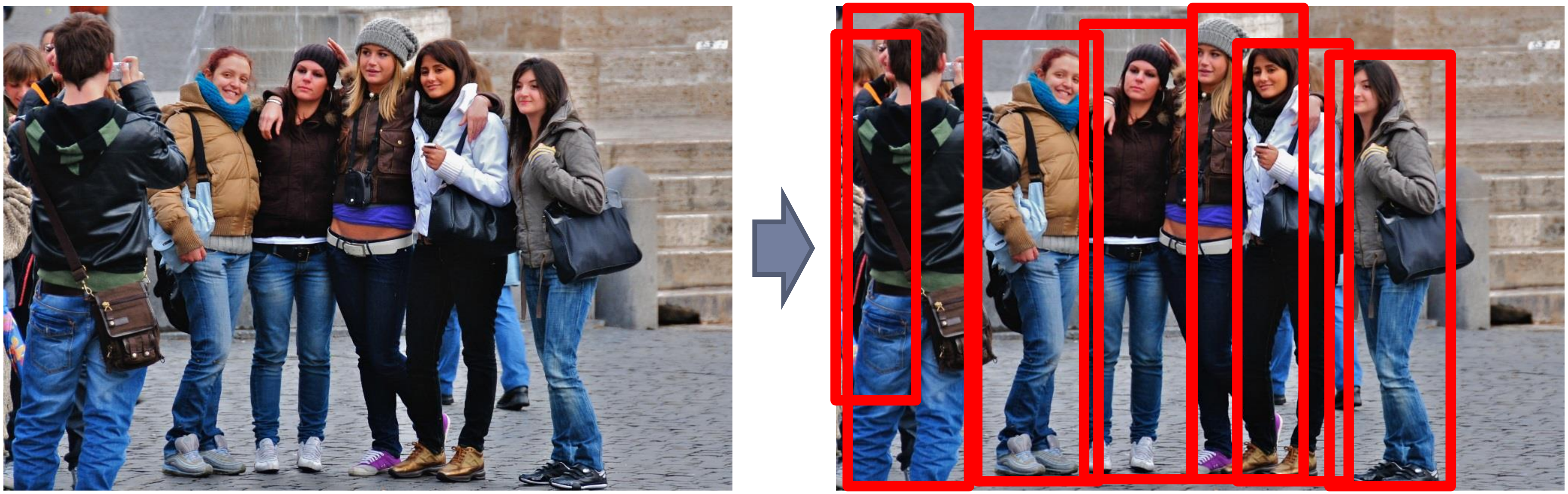
Visual Grounding: Task

The two girls in hats in the middle



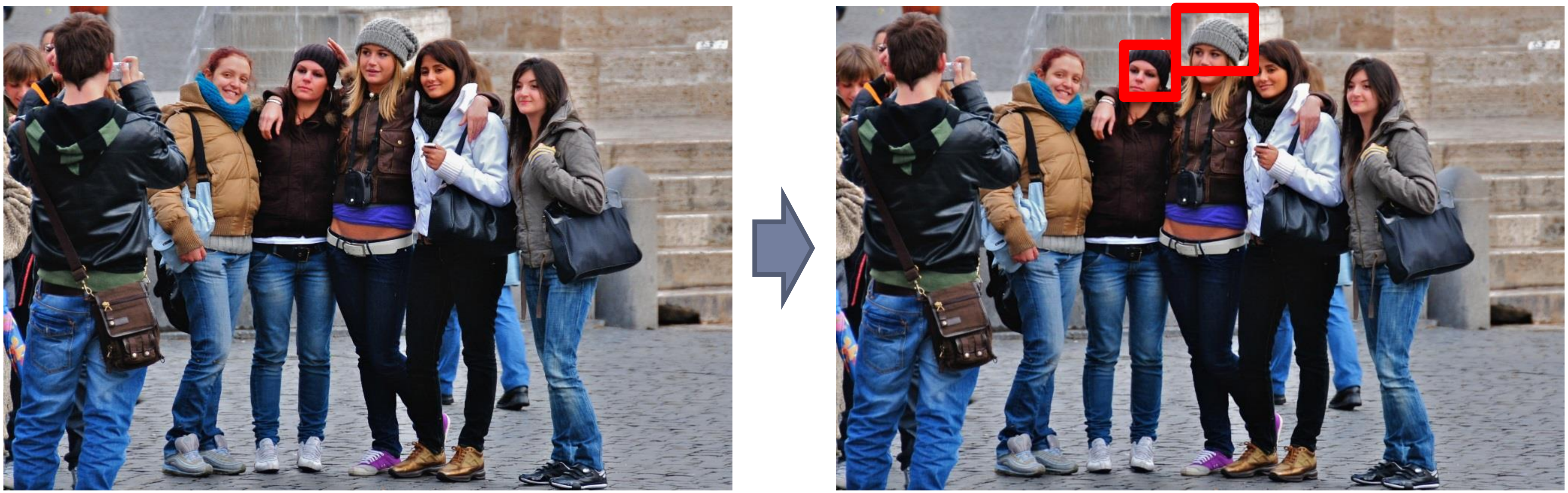
Visual Grounding: Task

Person detector



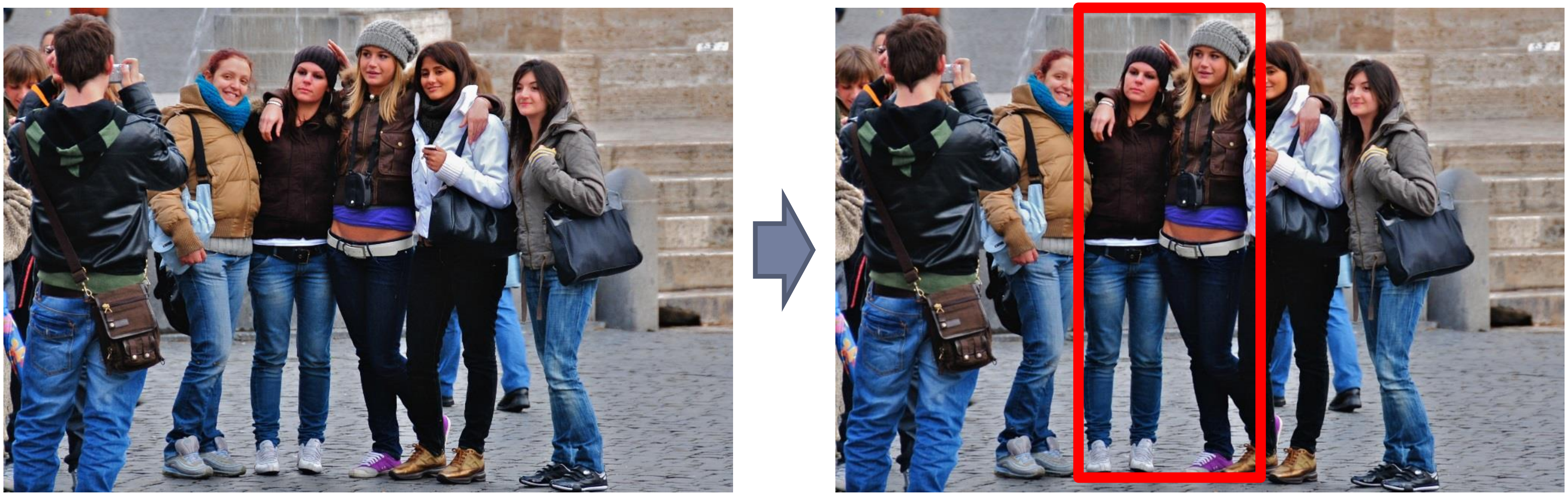
Visual Grounding: Task

Hat detector



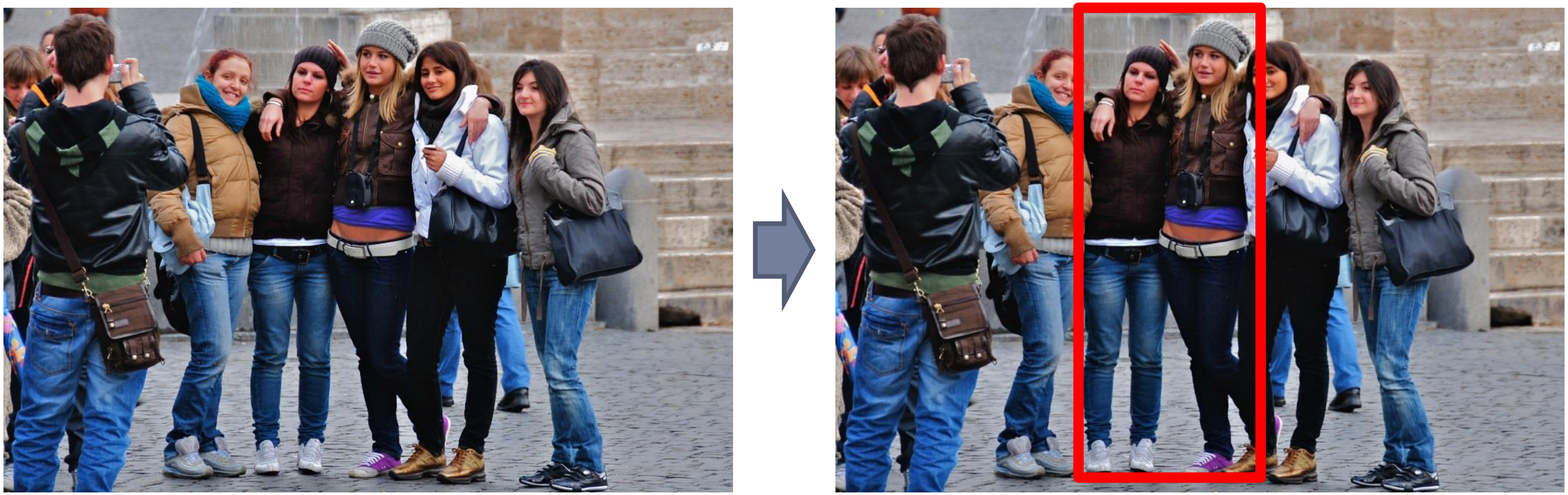
Visual Grounding: Task

The two girls in hats in the middle



Visual Grounding: Task

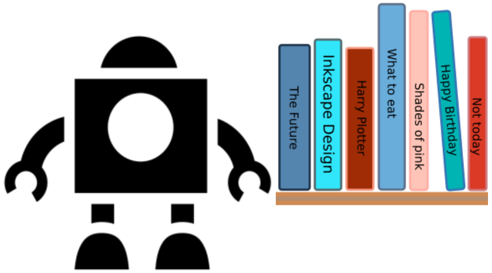
The two girls in hats in the middle



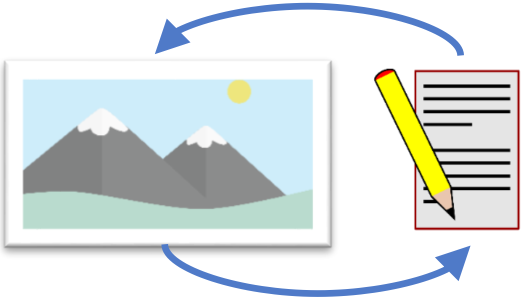
Visual Grounding: localize multi-word natural language expressions

Visual Grounding: Applications

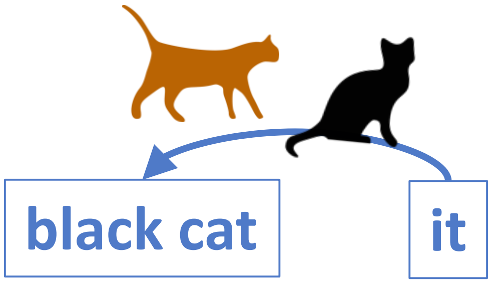
- Human-robot interaction
  - "Give me the middle blue book!"



- Supports other language-vision tasks
  - Captioning, VQA



- Text-image co-reference resolution
  - Resolve ambiguities

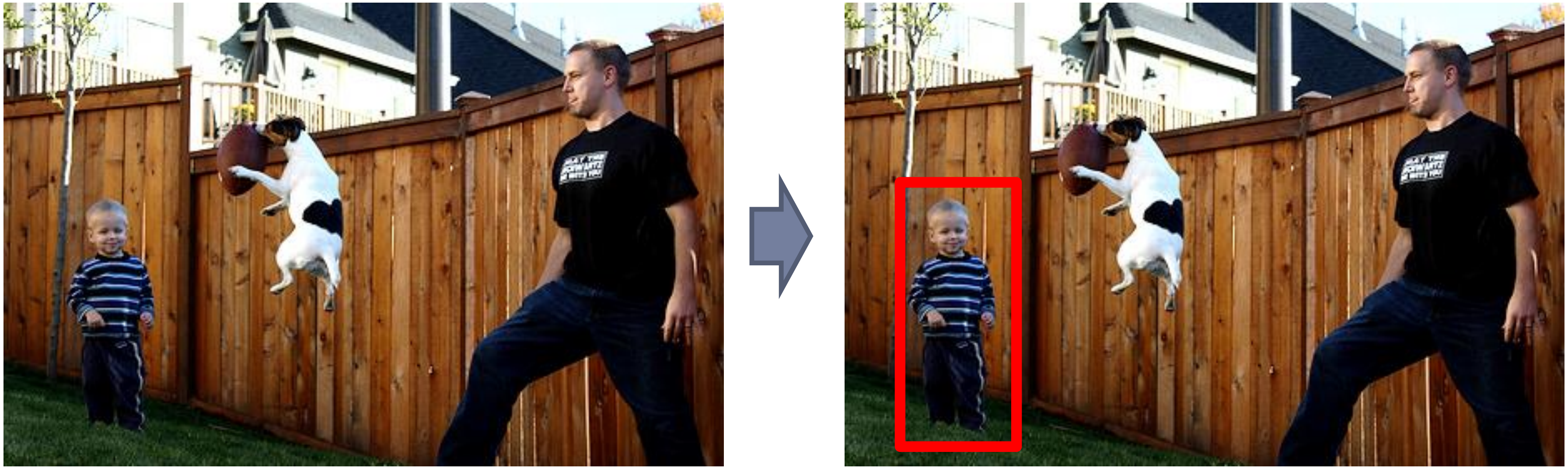


black cat

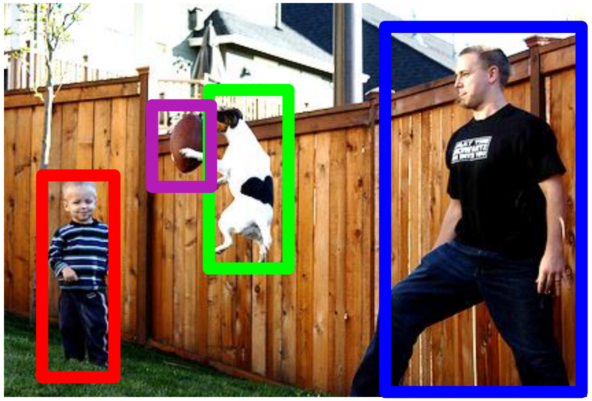
it

Visual Grounding: Test Time

a small boy



Visual Grounding: Training Time



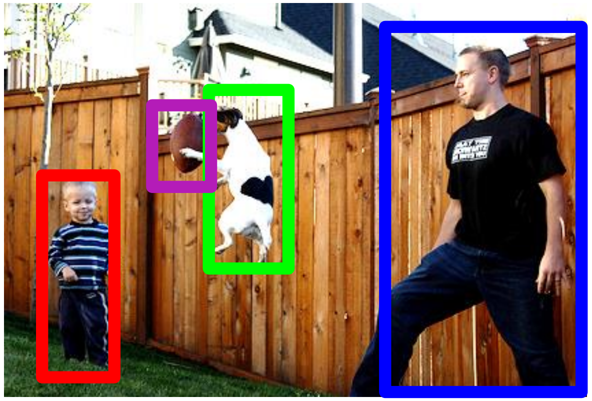
a small boy,
a man,
their small white dog,
a toy

Supervised

Visual Grounding: Training Time

a small boy,
a man,
their small white dog,
a toy

Unsupervised



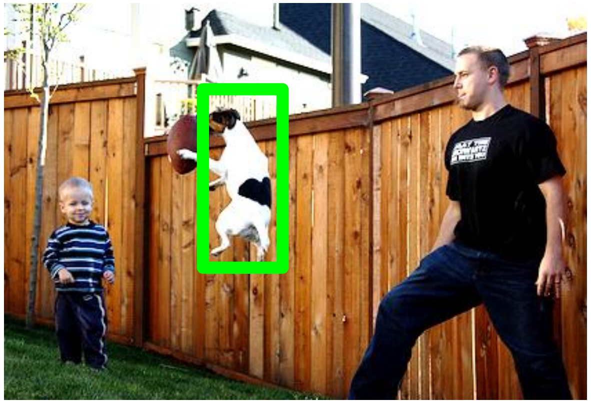
a small boy,
a man,
their small white dog,
a toy

Supervised

Visual Grounding: Training Time

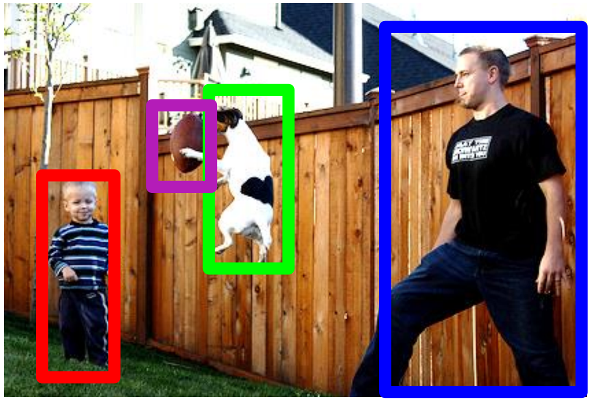
a small boy,
a man,
their small white dog,
a toy

Unsupervised



a small boy,
a man,
their small white dog,
a toy

Semi-supervised



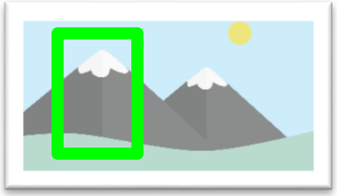
a small boy,
a man,
their small white dog,
a toy

Supervised

Visual Grounding: Training Time

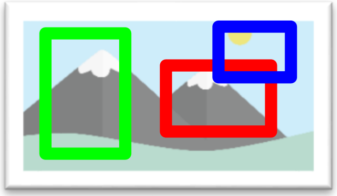
A B C

Unsupervised



A B C

Semi-supervised



A B C

Supervised

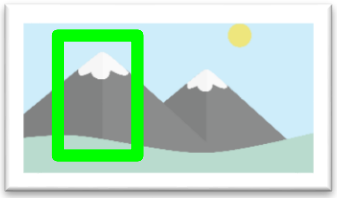
Visual Grounding: Related work

Karpathy NIPS’14
Karpathy CVPR’15

Plummer ICCV’15
Wang CVPR’16
Hu CVPR’16
Mao CVPR’16

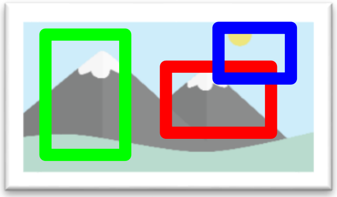
A B C

Unsupervised



A B C

Semi-supervised



A B C

Supervised

Visual Grounding: Related work

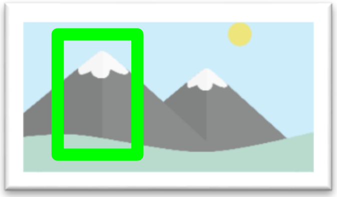
This work

Karpathy NIPS’14
Karpathy CVPR’15

A B C

Unsupervised

This work

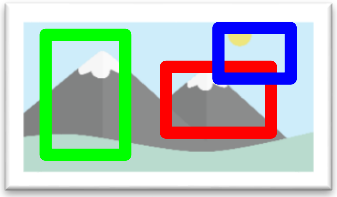


A B C

Semi-supervised

This work

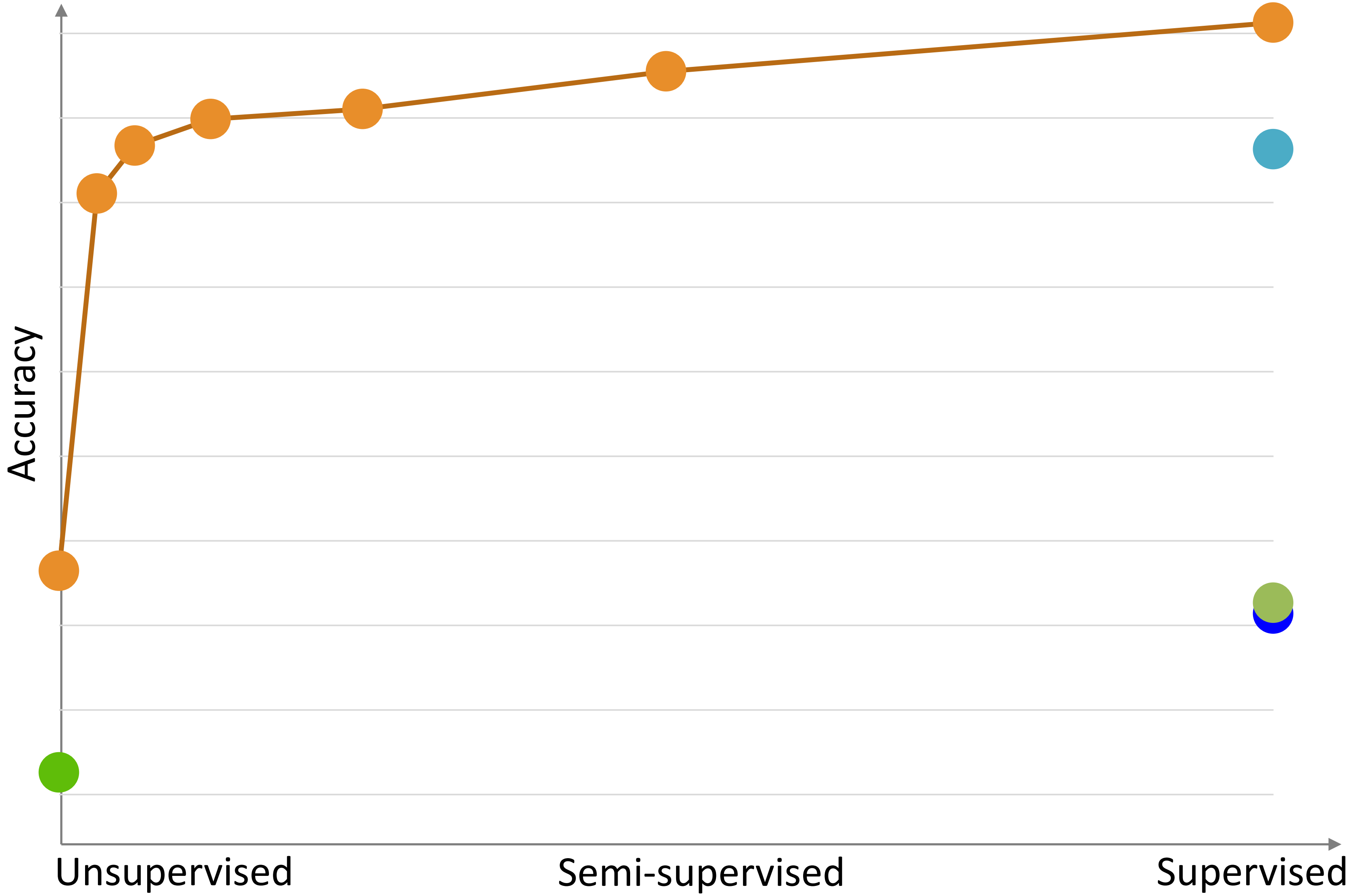
Plummer ICCV’15
Wang CVPR’16
Hu CVPR’16
Mao CVPR’16



A B C

Supervised

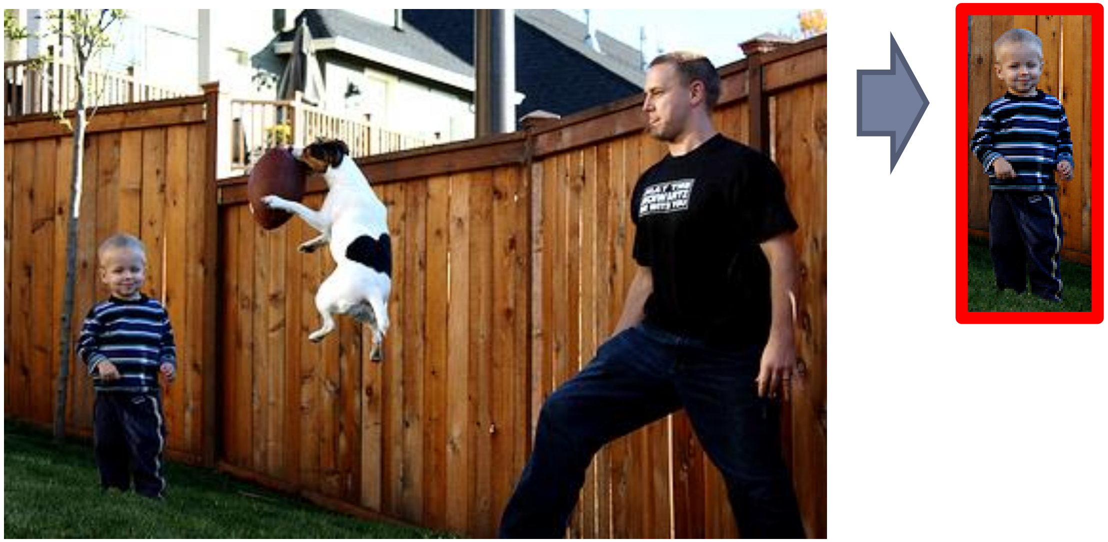
Main Result



Grounding approach

Grounding approach

a small boy

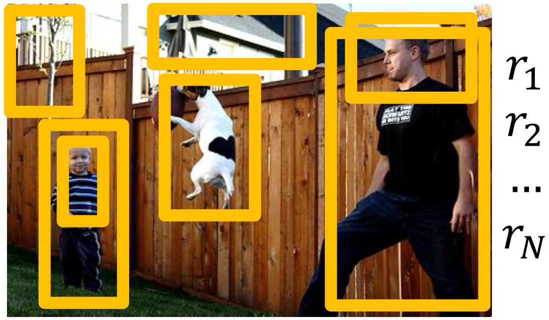


GroundeR: Grounding by Reconstruction

a small boy

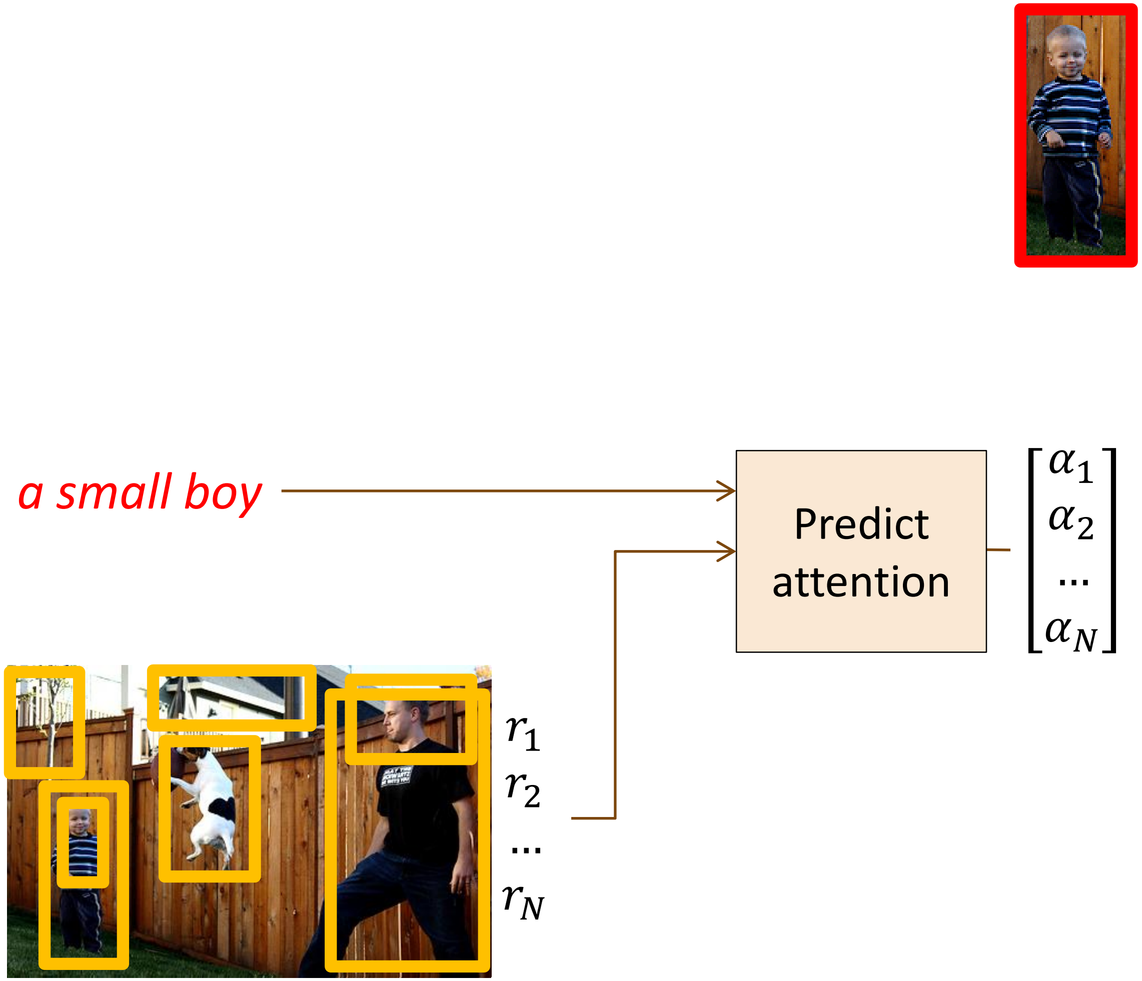
GroundeR: Grounding by Reconstruction

a small boy



Bounding box proposals

GroundeR: Grounding by Reconstruction



a small boy

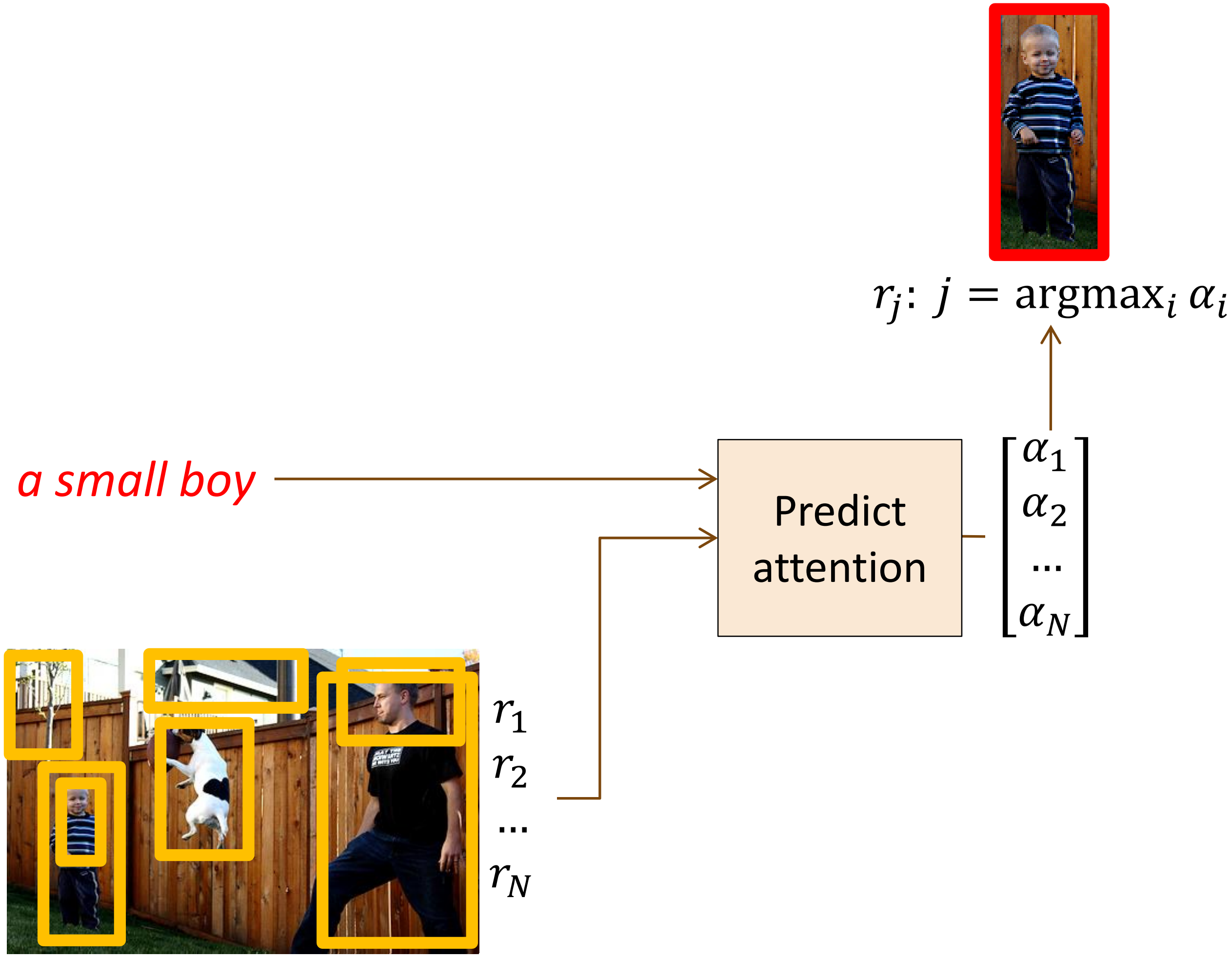
Predict attention

$$\begin{bmatrix}\alpha_1\\ \alpha_2\\ \cdots\\ \alpha_N\end{bmatrix}$$

r_1
r_2
$\ldots$
r_N

Bounding box proposals

GroundeR: Grounding by Reconstruction



$r_j\colon\ j = \mathrm{argmax}_i\ \alpha_i$

a small boy

Predict attention

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \cdots \\ \alpha_N \end{bmatrix}$$

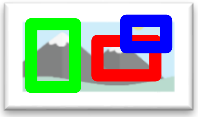
r_1
r_2
$\cdots$
r_N

Bounding box proposals

GroundeR: Grounding by Reconstruction

$r_j\colon\ j = \operatorname{argmax}_i \alpha_i$

Supervision

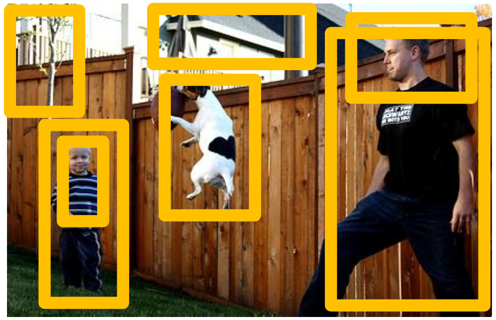


A B C

a small boy

Predict attention

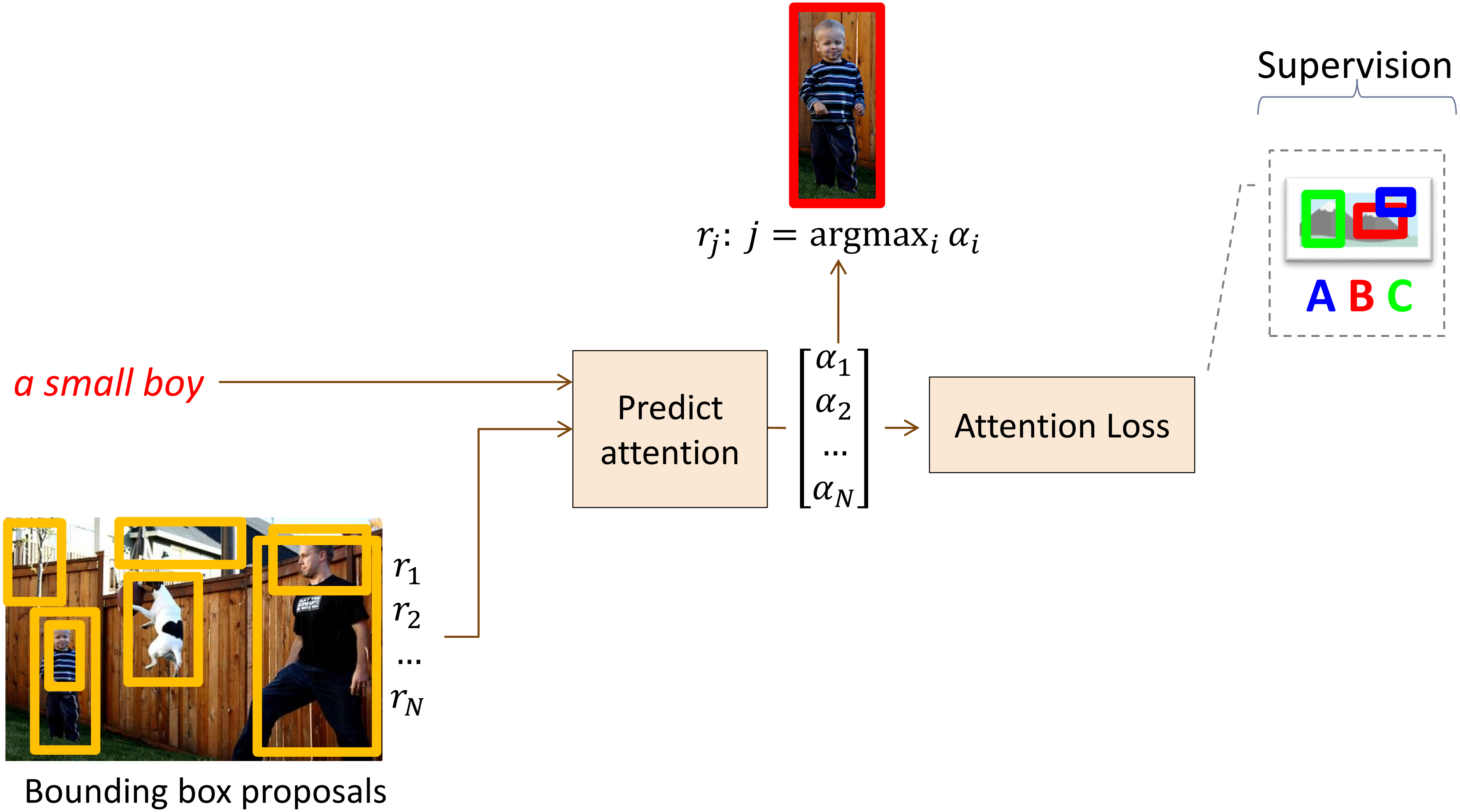
$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \cdots \\ \alpha_N \end{bmatrix}$$



r_1
r_2
$\cdots$
r_N

Bounding box proposals

GroundeR: Grounding by Reconstruction



Supervision

A B C

r_j: $j = \text{argmax}_i \, \alpha_i$

a small boy

Predict attention

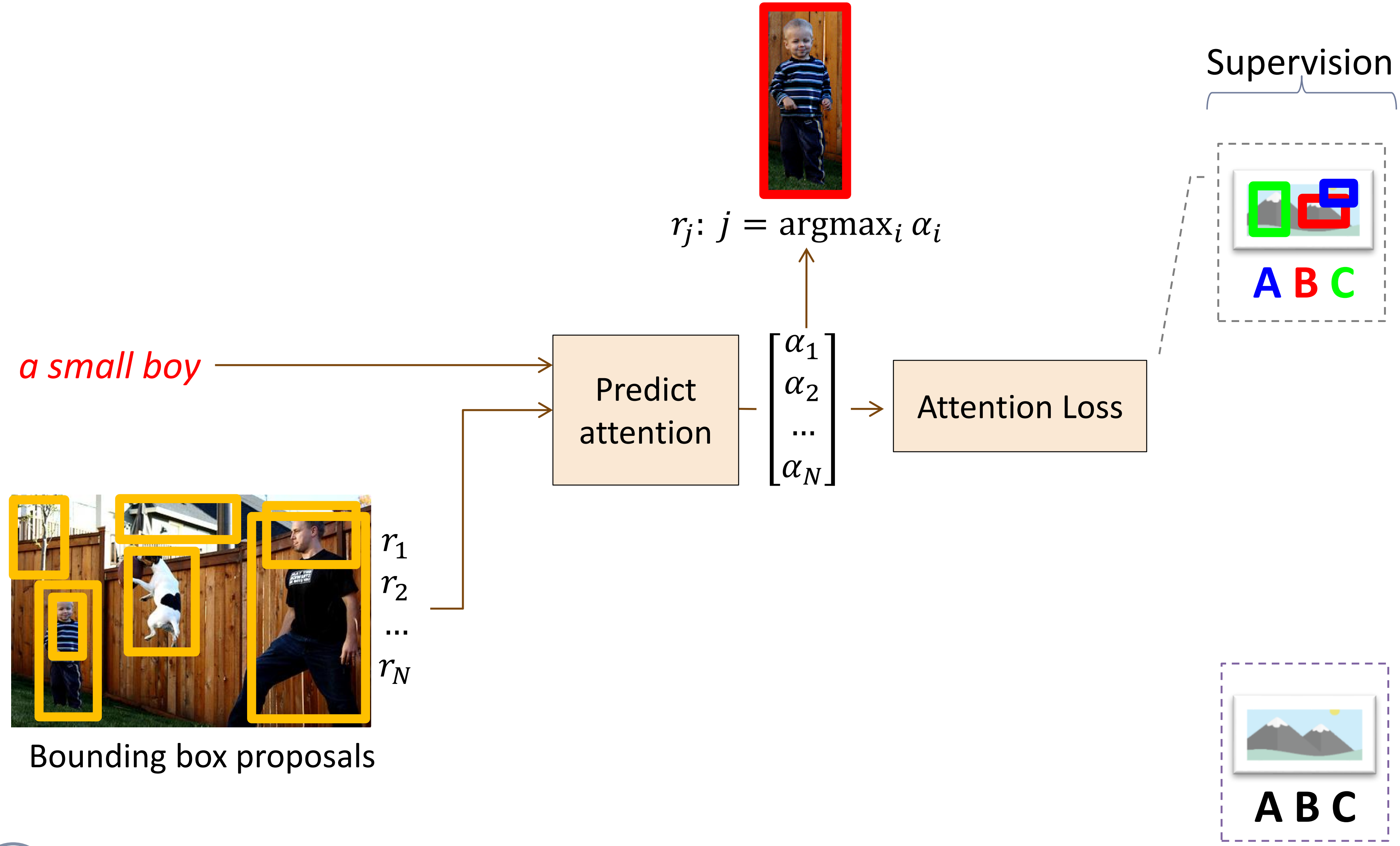
$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \cdots \\ \alpha_N \end{bmatrix}$$

Attention Loss

r_1
r_2
$\cdots$
r_N

Bounding box proposals

GroundeR: Grounding by Reconstruction



Supervision

A B C

r_j: $j = \text{argmax}_i\, \alpha_i$

a small boy

Predict attention

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \cdots \\ \alpha_N \end{bmatrix}$$

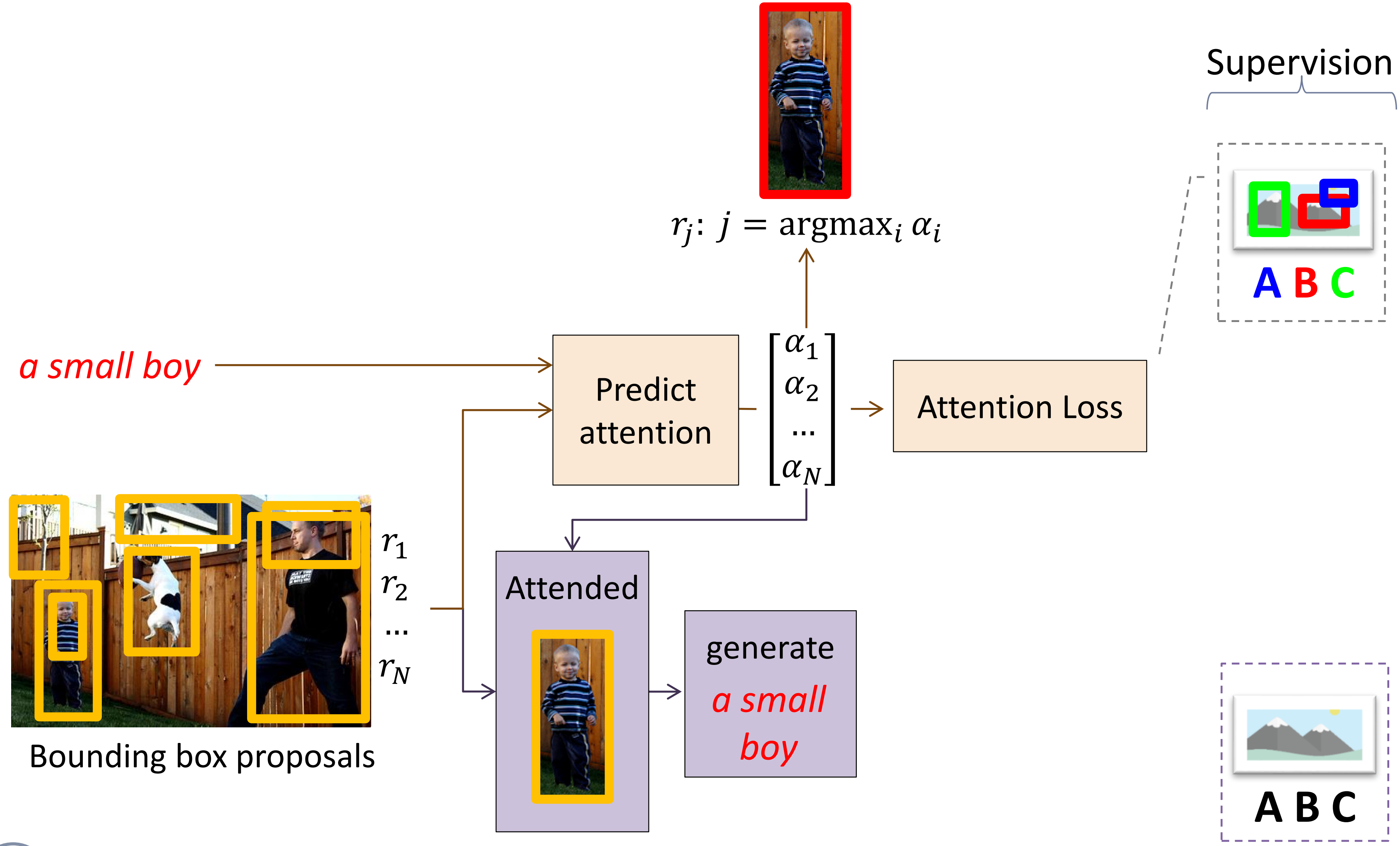
Attention Loss

r_1
r_2
$\cdots$
r_N

Bounding box proposals

A B C

GroundeR: Grounding by Reconstruction



Supervision

A B C

r_j: $j = \mathrm{argmax}_i\ \alpha_i$

a small boy

Predict attention

$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \cdots \\ \alpha_N \end{bmatrix}$

Attention Loss

r_1

r_2

…

r_N

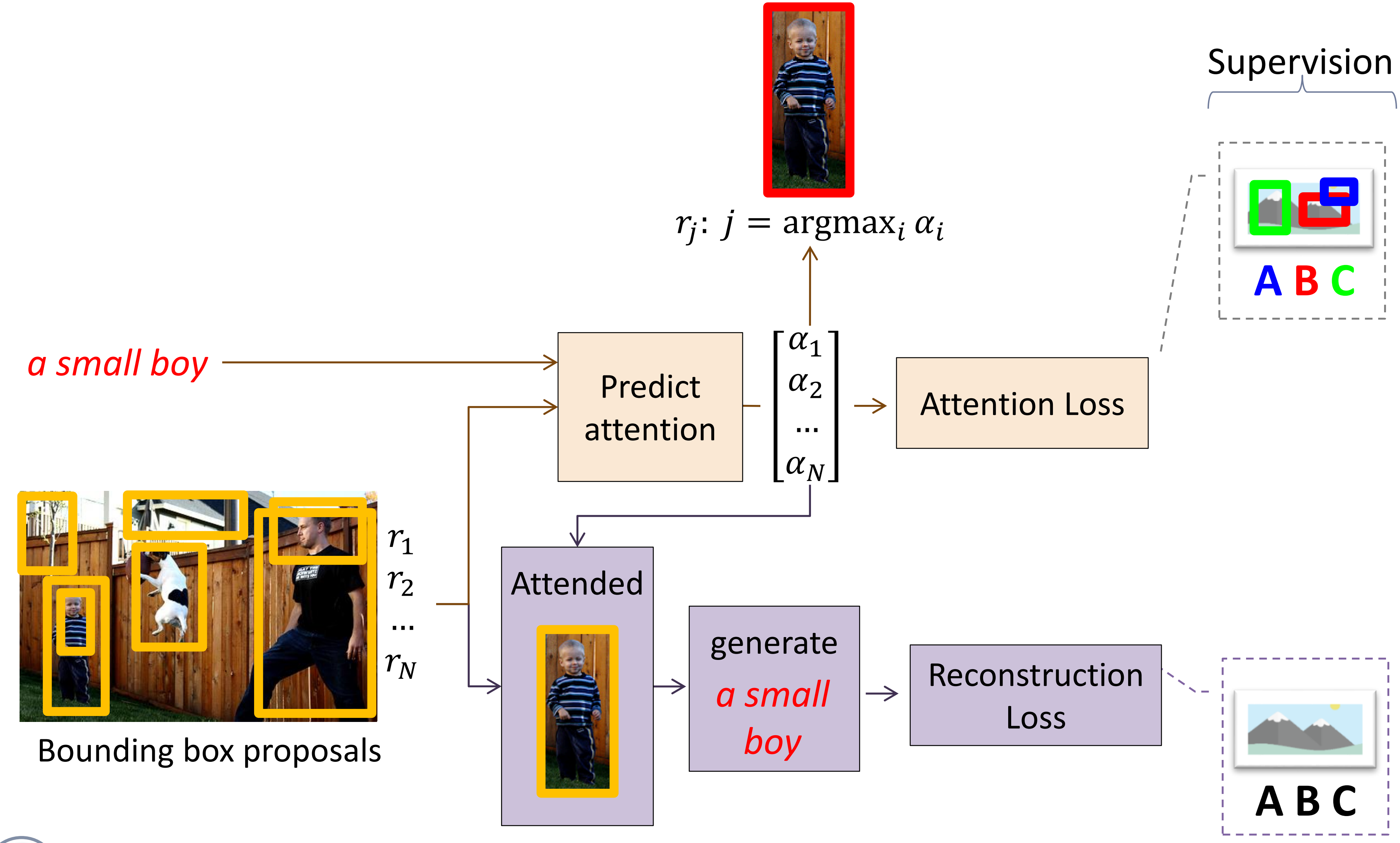
Bounding box proposals

Attended

generate *a small boy*

A B C

GroundeR: Grounding by Reconstruction



Supervision

A B C

r_j: $j = \text{argmax}_i\ \alpha_i$

a small boy

Predict attention

$$\begin{bmatrix}\alpha_1\\ \alpha_2\\ \cdots\\ \alpha_N\end{bmatrix}$$

Attention Loss

r_1
r_2
$\ldots$
r_N

Bounding box proposals

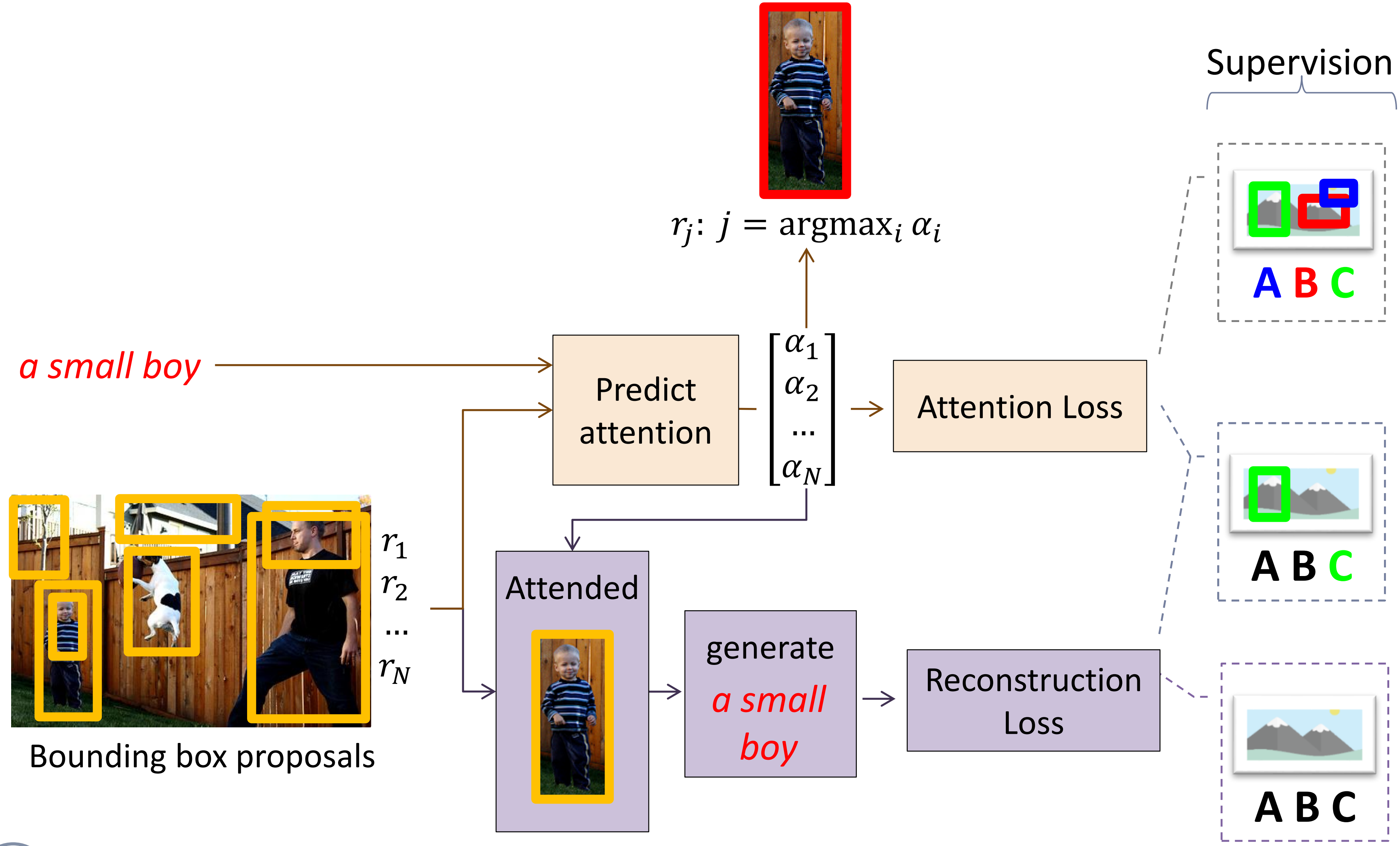
Attended

generate *a small boy*

Reconstruction Loss

A B C

GroundeR: Grounding by Reconstruction



Supervision

A B C

A B C

A B C

r_j: $j = \text{argmax}_i\, \alpha_i$

a small boy

Predict attention

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \cdots \\ \alpha_N \end{bmatrix}$$

Attention Loss

r_1

r_2

…

r_N

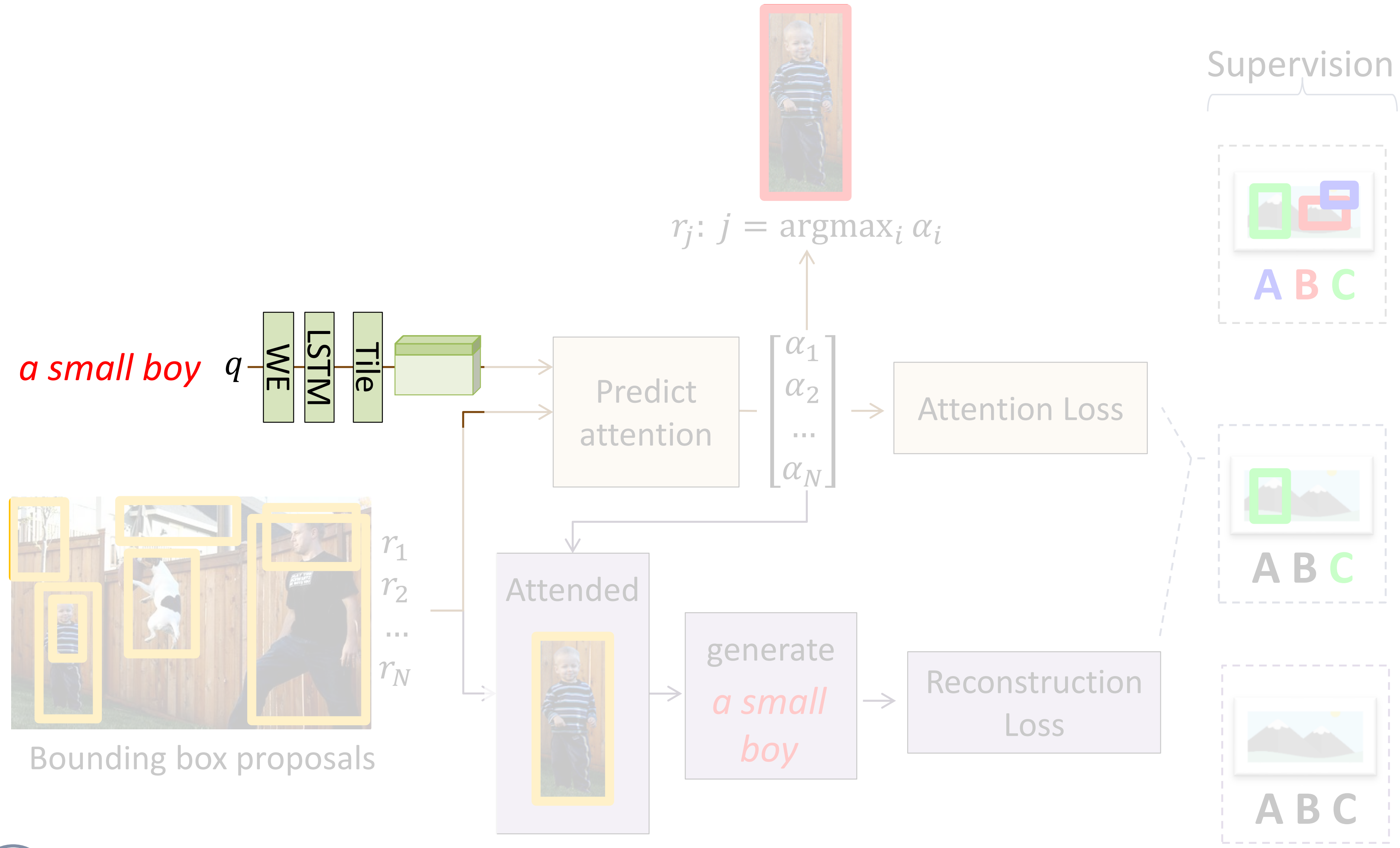
Bounding box proposals

Attended

generate *a small boy*

Reconstruction Loss

GroundeR: Grounding by Reconstruction



Supervision

A B C

A B C

A B C

r_j: $j = \text{argmax}_i\ \alpha_i$

a small boy q WE LSTM Tile

Predict attention

$\begin{bmatrix}\alpha_1\\ \alpha_2\\ \dots\\ \alpha_N\end{bmatrix}$

Attention Loss

r_1 r_2 … r_N

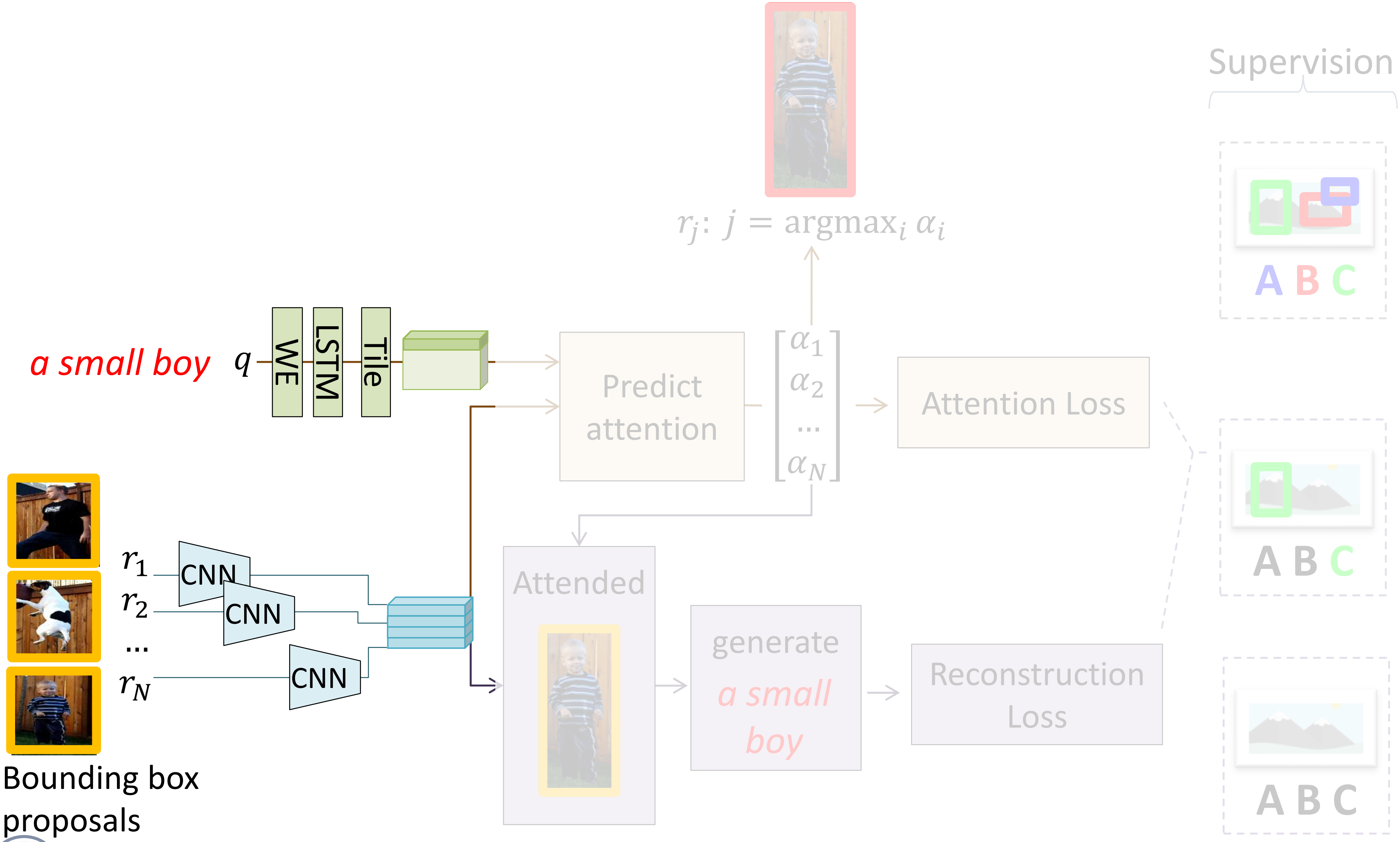
Bounding box proposals

Attended

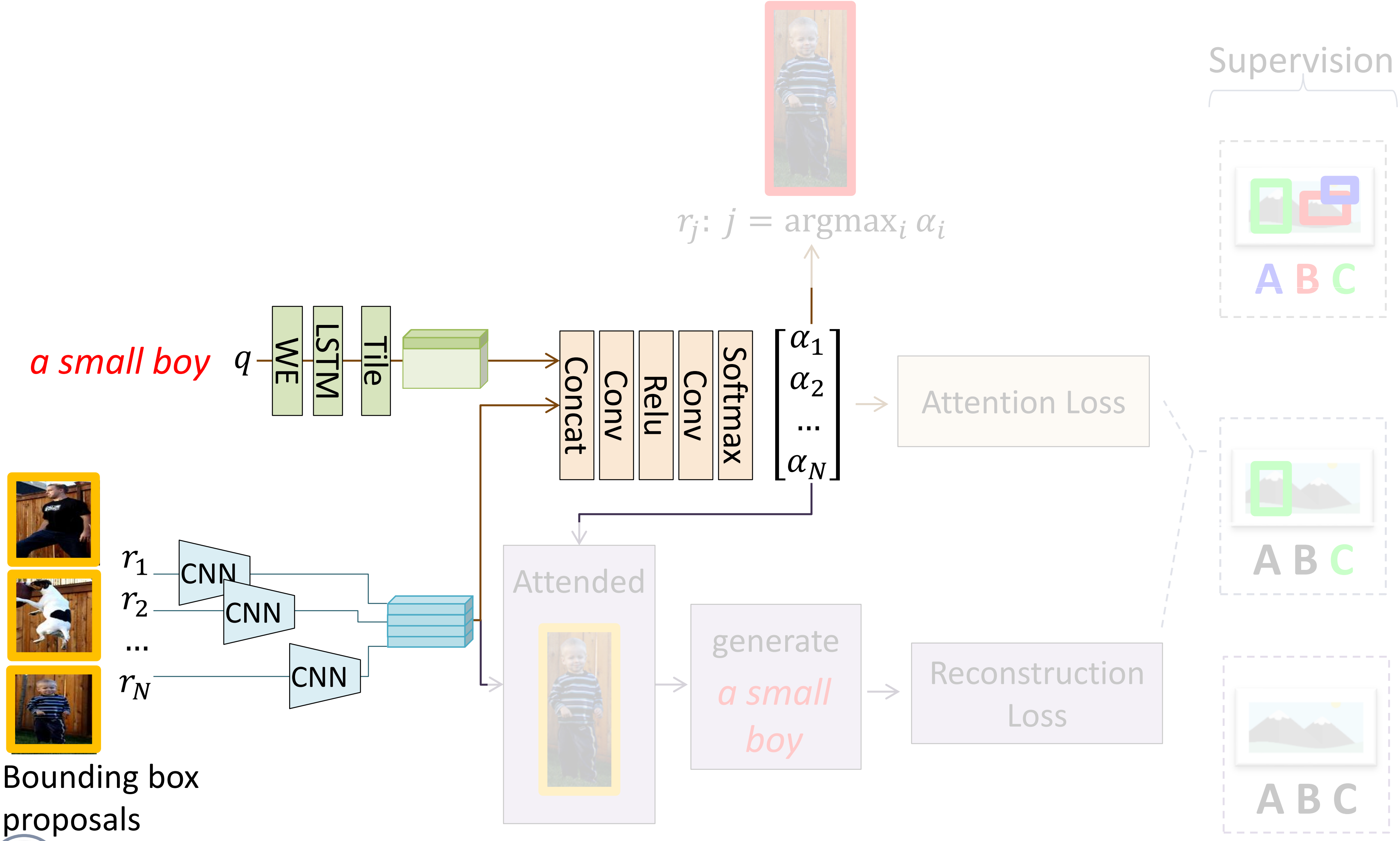
generate *a small boy*

Reconstruction Loss

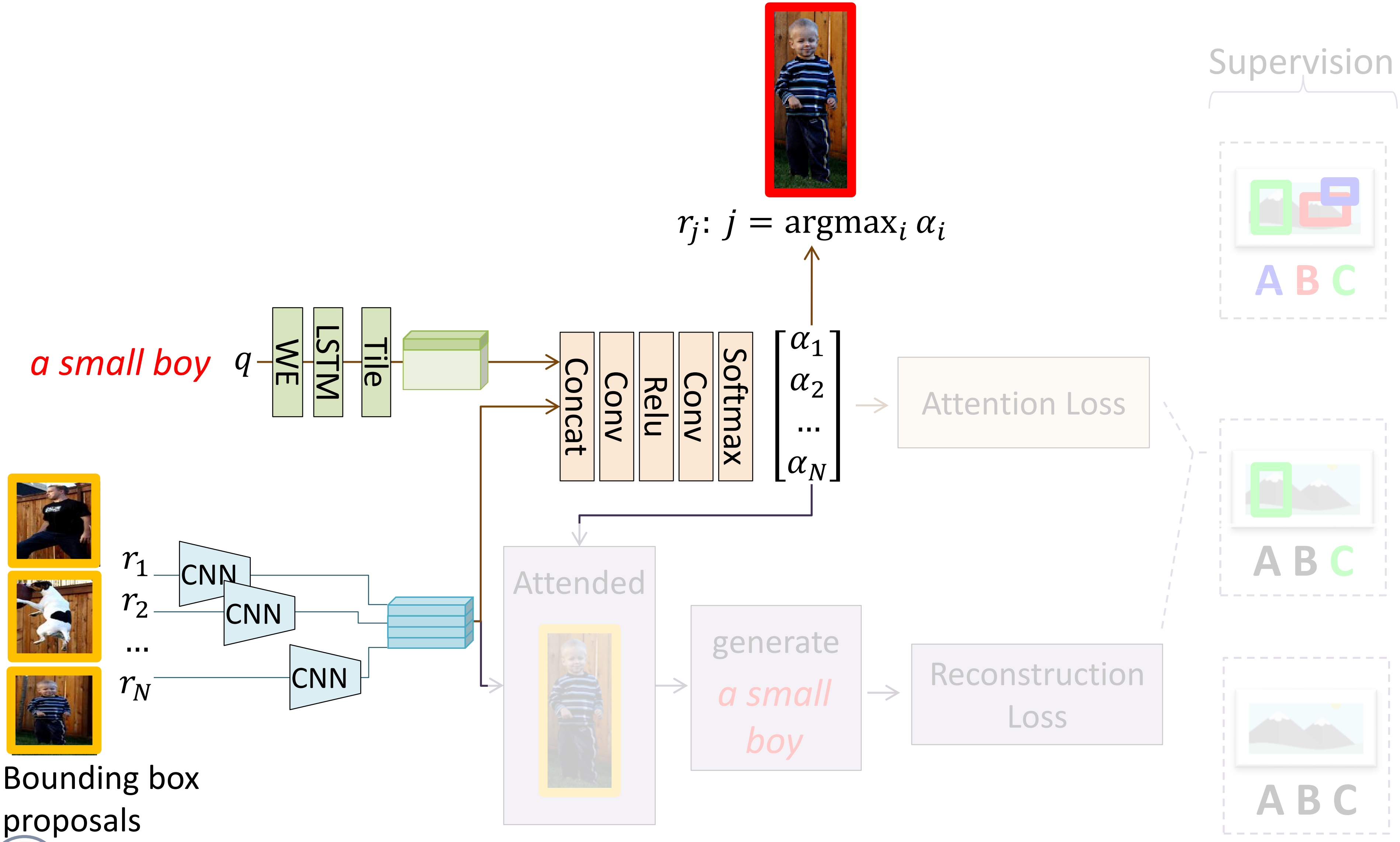
GroundeR: Grounding by Reconstruction



GroundeR: Grounding by Reconstruction



GroundeR: Grounding by Reconstruction



r_j: $j = \text{argmax}_i\, \alpha_i$

Supervision

A B C

A B C

A B C

a small boy q

WE

LSTM

Tile

Concat

Conv

Relu

Conv

Softmax

$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \ldots \\ \alpha_N \end{bmatrix}$

Attention Loss

r_1

r_2

…

r_N

CNN

CNN

CNN

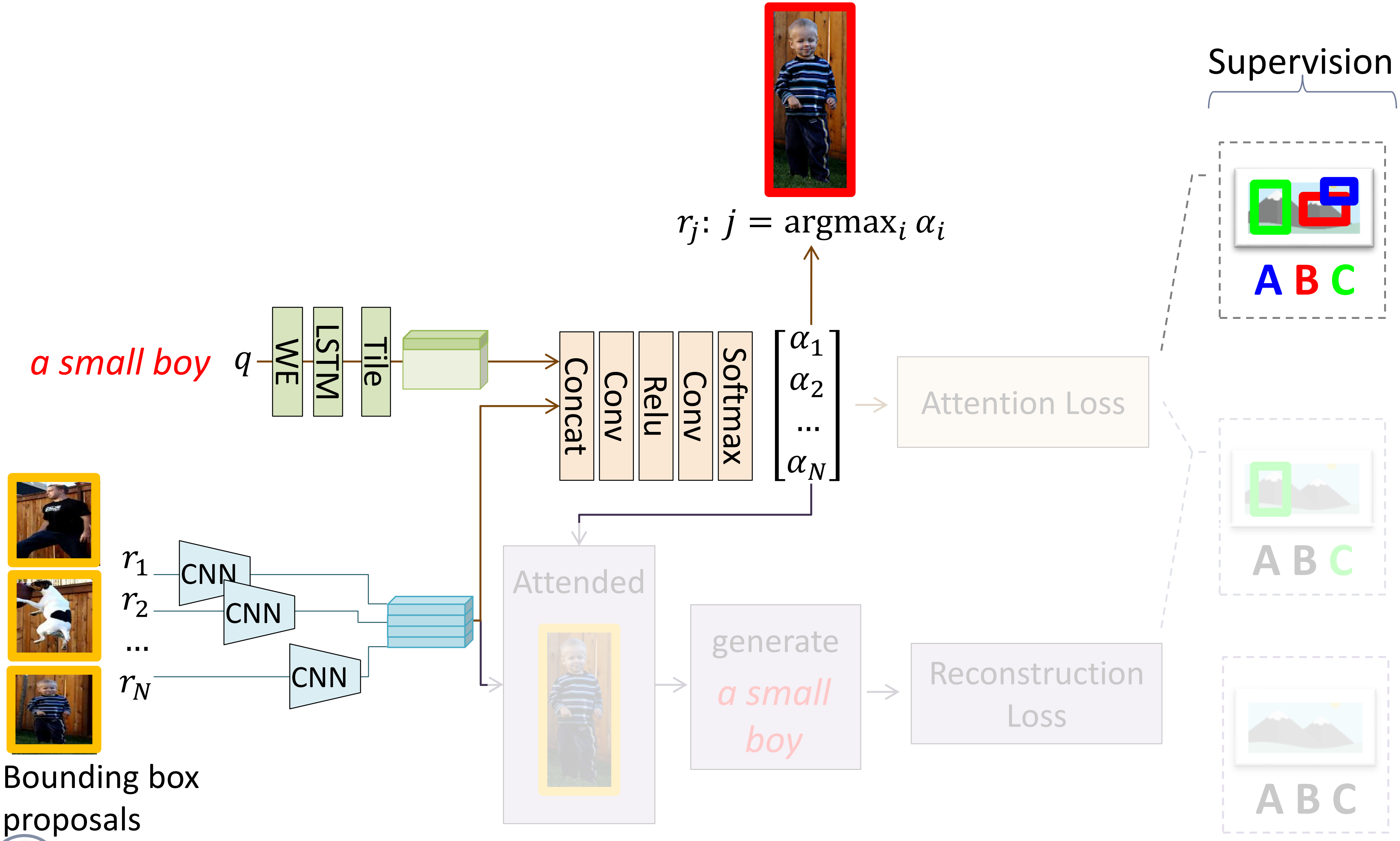
Attended

generate *a small boy*

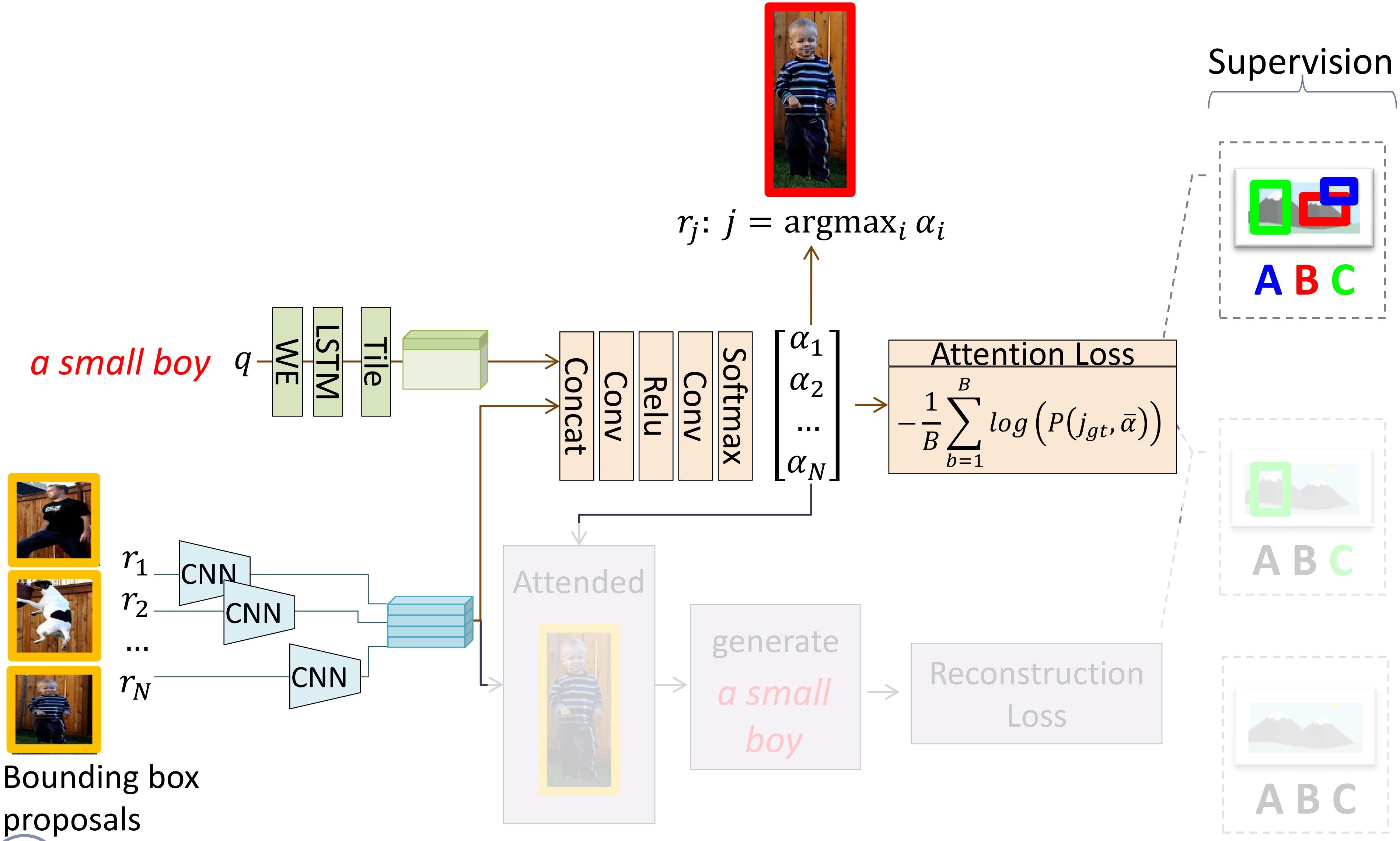
Reconstruction Loss

Bounding box proposals

GroundeR: Grounding by Reconstruction



GroundeR: Grounding by Reconstruction



Supervision

A B C

A B C

A B C

r_j: $j = \text{argmax}_i \, \alpha_i$

a small boy q

WE

LSTM

Tile

Concat

Conv

Relu

Conv

Softmax

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \dots \\ \alpha_N \end{bmatrix}$$

Attention Loss

$$-\frac{1}{B}\sum_{b=1}^{B} log\left(P\left(j_{gt},\bar{\alpha}\right)\right)$$

r_1 CNN

r_2 CNN

…

r_N CNN

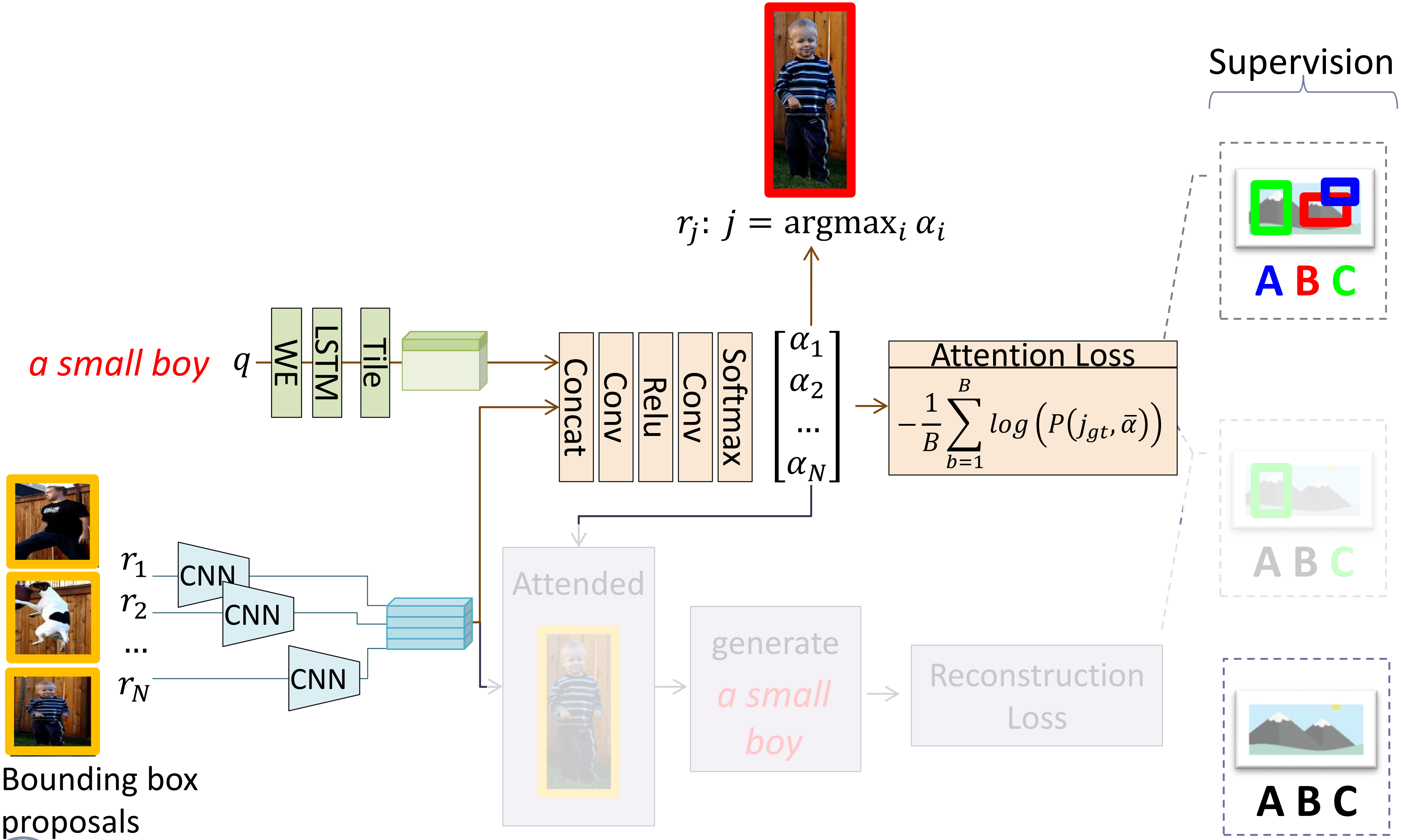
Bounding box proposals

Attended

generate *a small boy*

Reconstruction Loss

GroundeR: Grounding by Reconstruction



Supervision

A B C

a small boy q

WE

LSTM

Tile

Concat

Conv

Relu

Conv

Softmax

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \dots \\ \alpha_N \end{bmatrix}$$

$$r_j\text{: } j = \text{argmax}_i \, \alpha_i$$

Attention Loss

$$-\frac{1}{B}\sum_{b=1}^{B} log\left(P\left(j_{gt}, \bar{\alpha}\right)\right)$$

A B C

A B C

r_1 CNN

r_2 CNN

…

r_N CNN

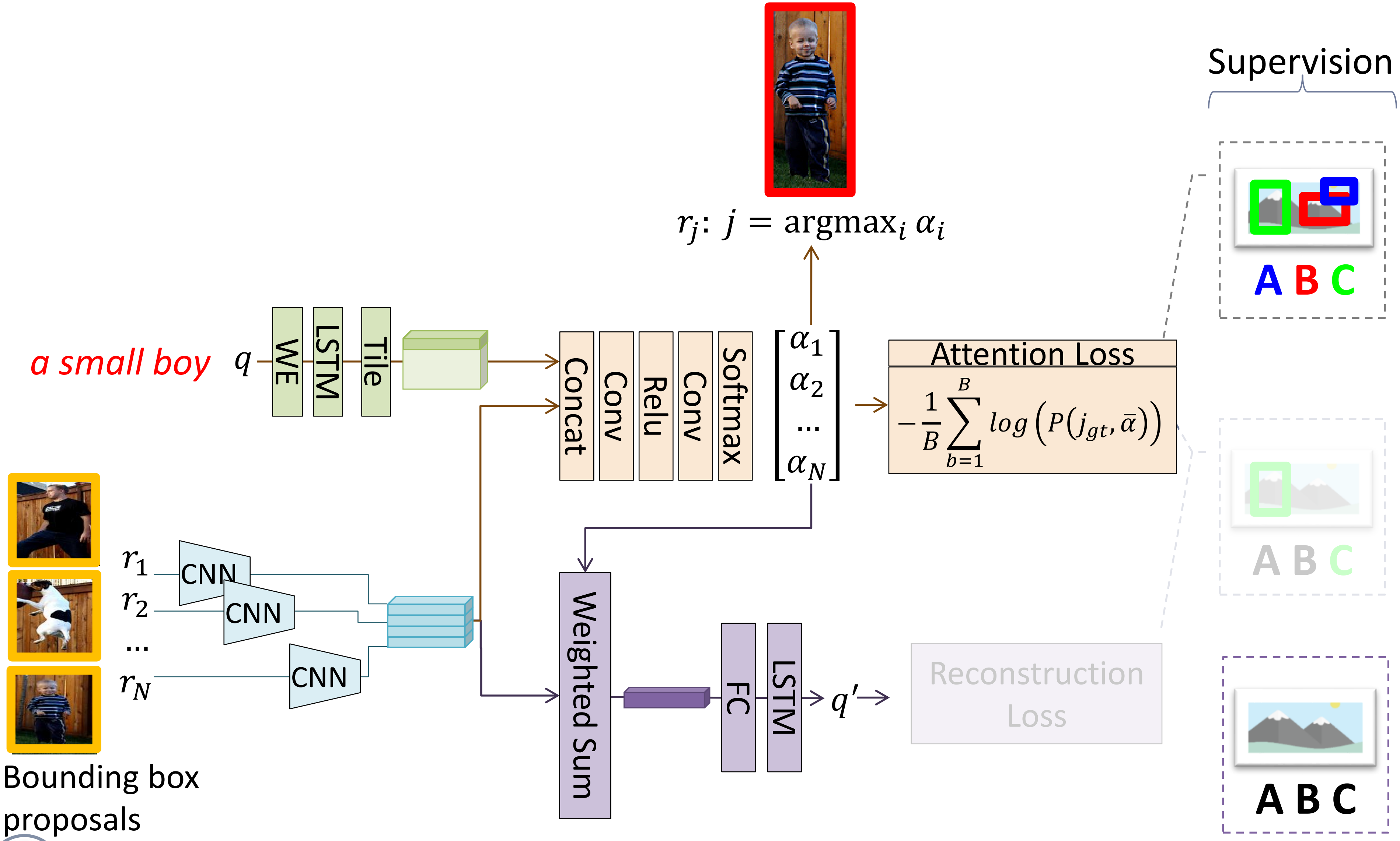
Bounding box proposals

Attended

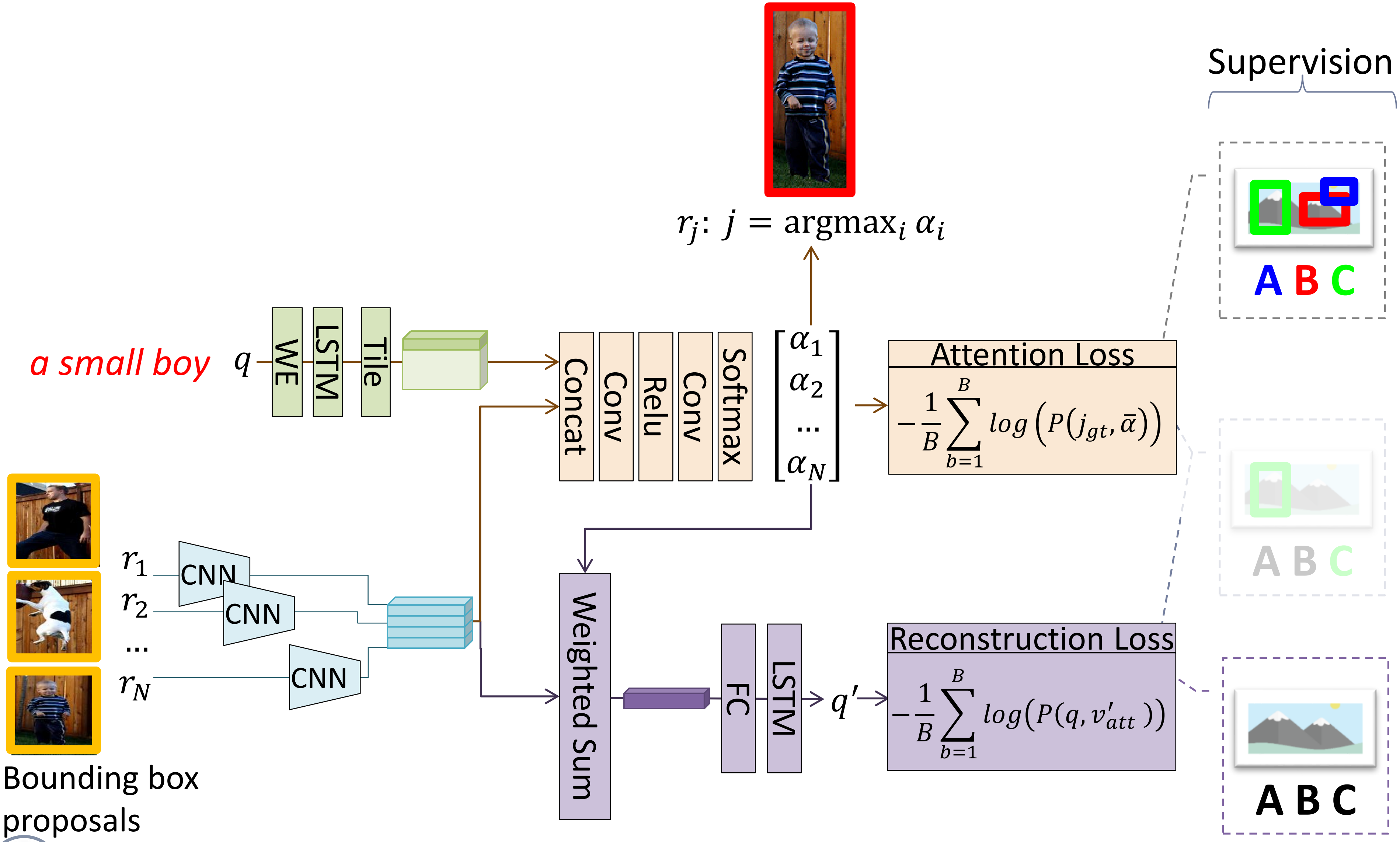
generate *a small boy*

Reconstruction Loss

GroundeR: Grounding by Reconstruction



GroundeR: Grounding by Reconstruction



Supervision

A B C

A B C

A B C

a small boy q

WE

LSTM

Tile

Concat

Conv

Relu

Conv

Softmax

r_j: $j = \text{argmax}_i \, \alpha_i$

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \dots \\ \alpha_N \end{bmatrix}$$

Attention Loss

$$-\frac{1}{B}\sum_{b=1}^{B} log\left(P\left(j_{gt}, \bar{\alpha}\right)\right)$$

r_1 CNN

r_2 CNN

…

r_N CNN

Bounding box proposals

Weighted Sum

FC

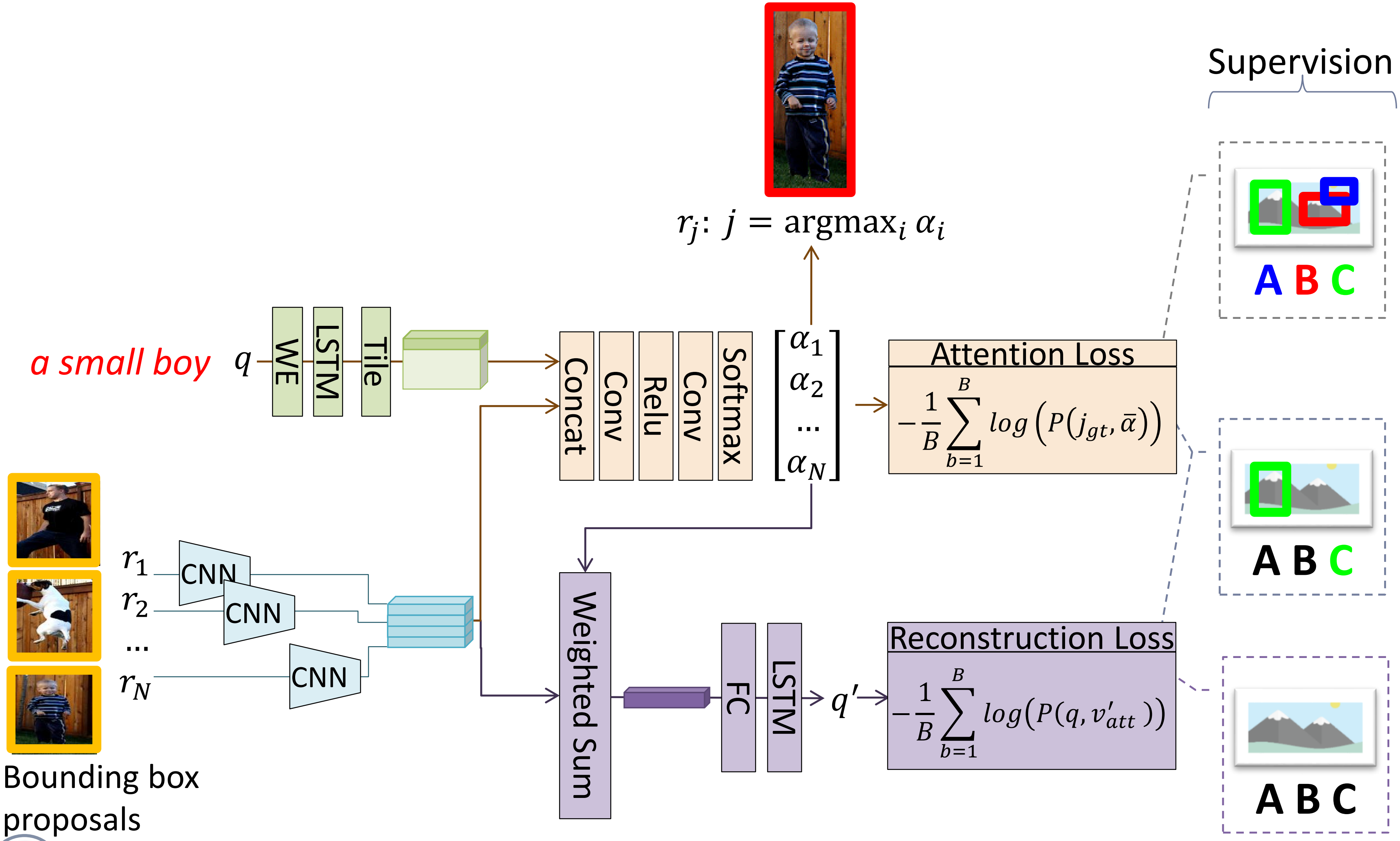
LSTM

q'

Reconstruction Loss

$$-\frac{1}{B}\sum_{b=1}^{B} log\left(P(q, v'_{att})\right)$$

GroundeR: Grounding by Reconstruction



Attention Loss

$$-\frac{1}{B}\sum_{b=1}^{B} log\left(P\left(j_{gt}, \bar{\alpha}\right)\right)$$

Reconstruction Loss

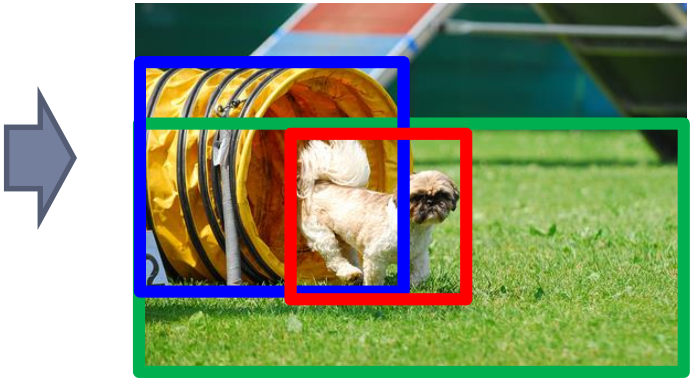
$$-\frac{1}{B}\sum_{b=1}^{B} log\left(P\left(q, v'_{att}\right)\right)$$

Experimental evaluation

Datasets and Experimental setup

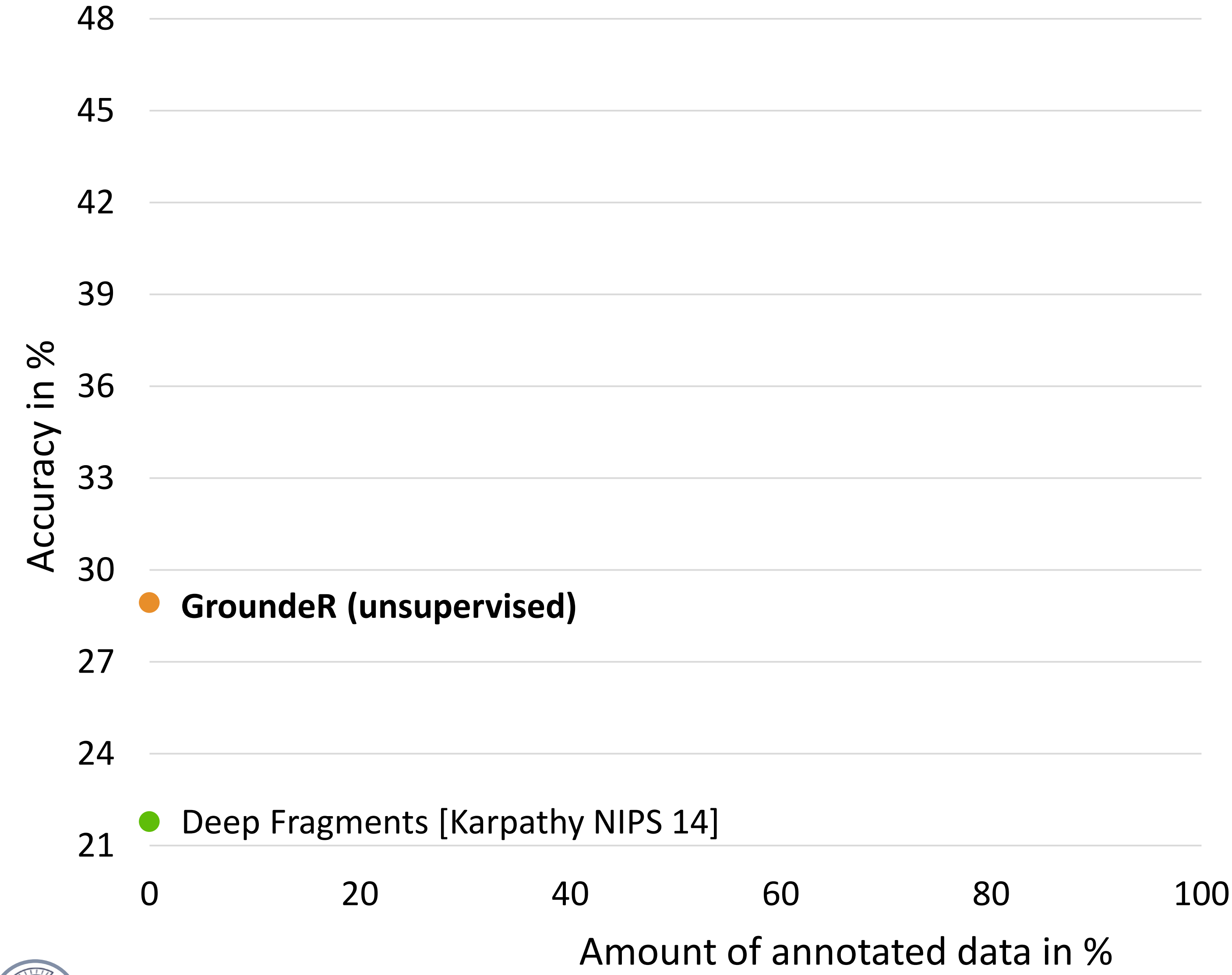
- Flickr30k Entities [Plummer ICCV'15]
  - 275k bounding boxes & noun phrases

A little brown and white dog emerged from a yellow collapsible toy tunnel onto the lawn.

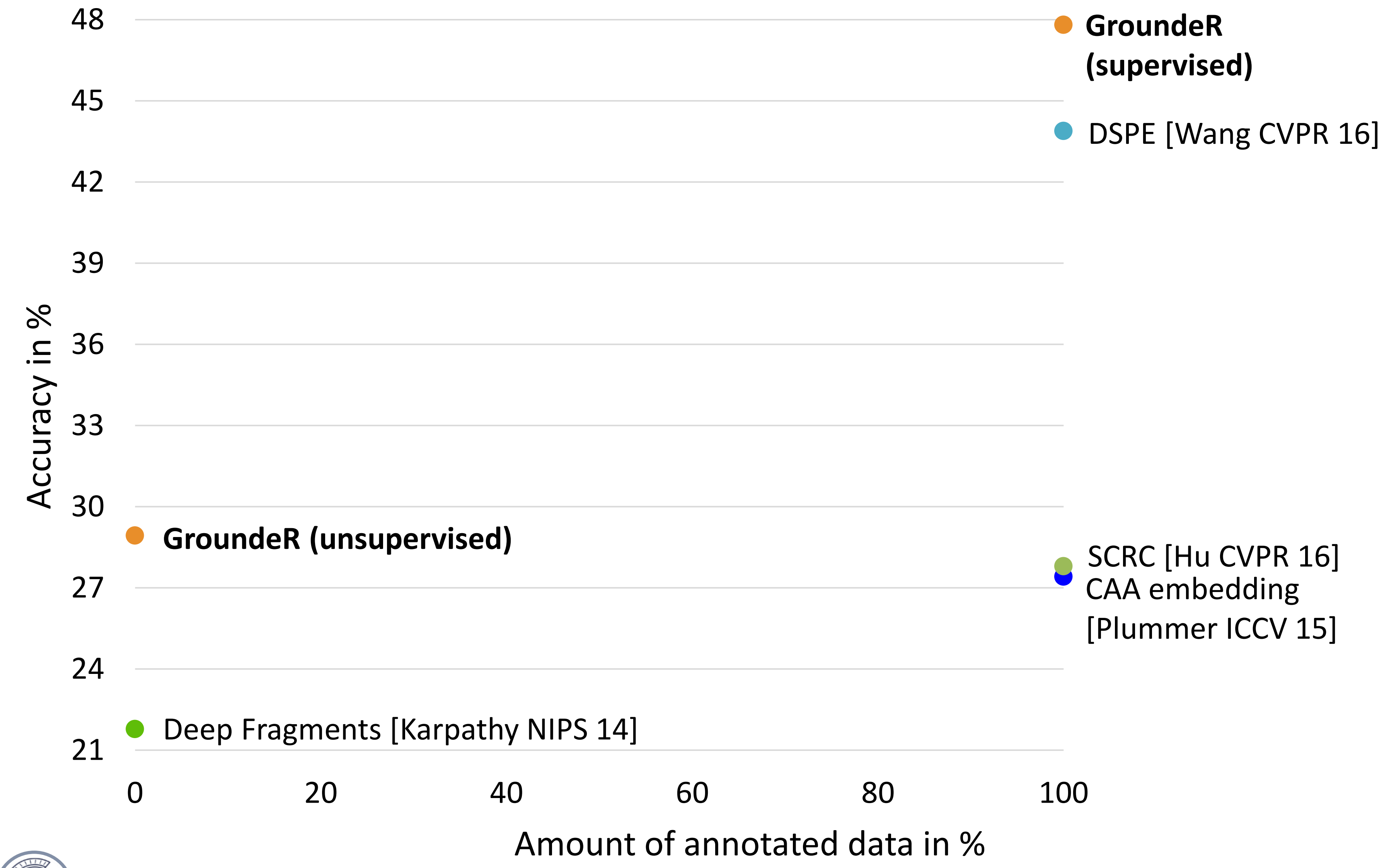


- Experimental setup
  - 100 object proposals
  - Fast R-CNN [Girshick ICCV'15]
- Accuracy
  - % phrases: IOU(predicted box, ground-truth box) >= 0.5

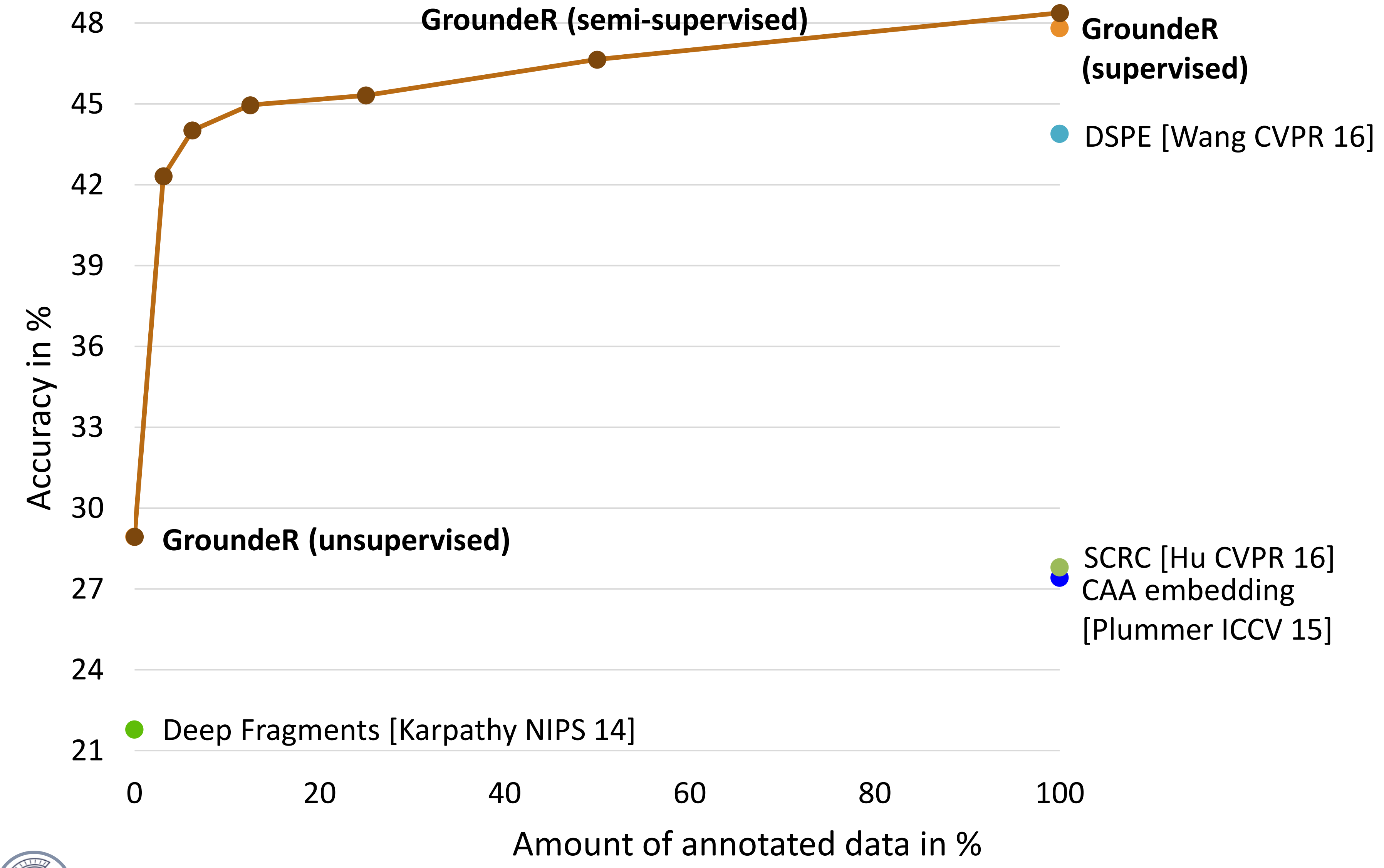
Evaluation: GroundeR on Flickr30k Entities



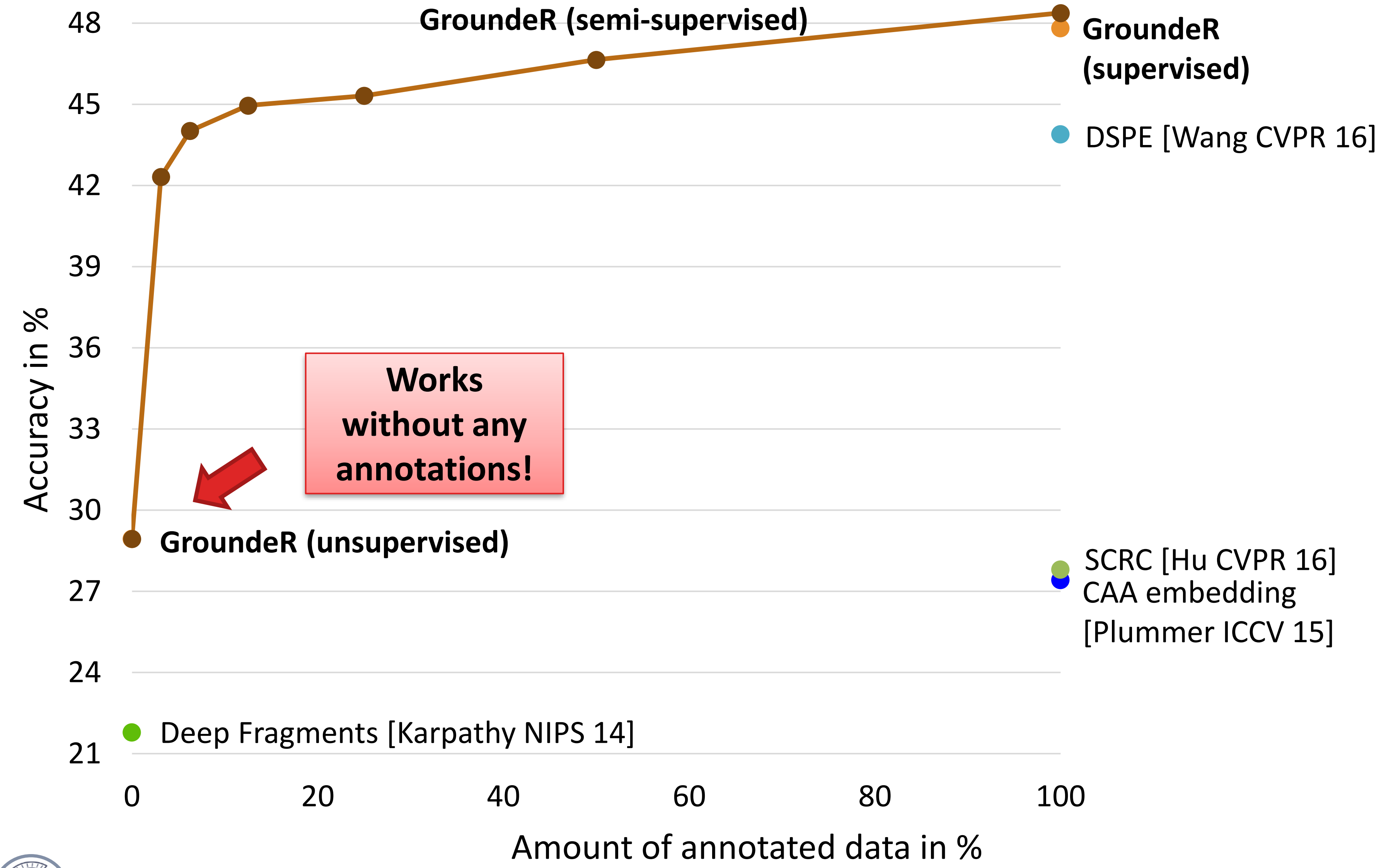
Evaluation: GroundeR on Flickr30k Entities



Evaluation: GroundeR on Flickr30k Entities

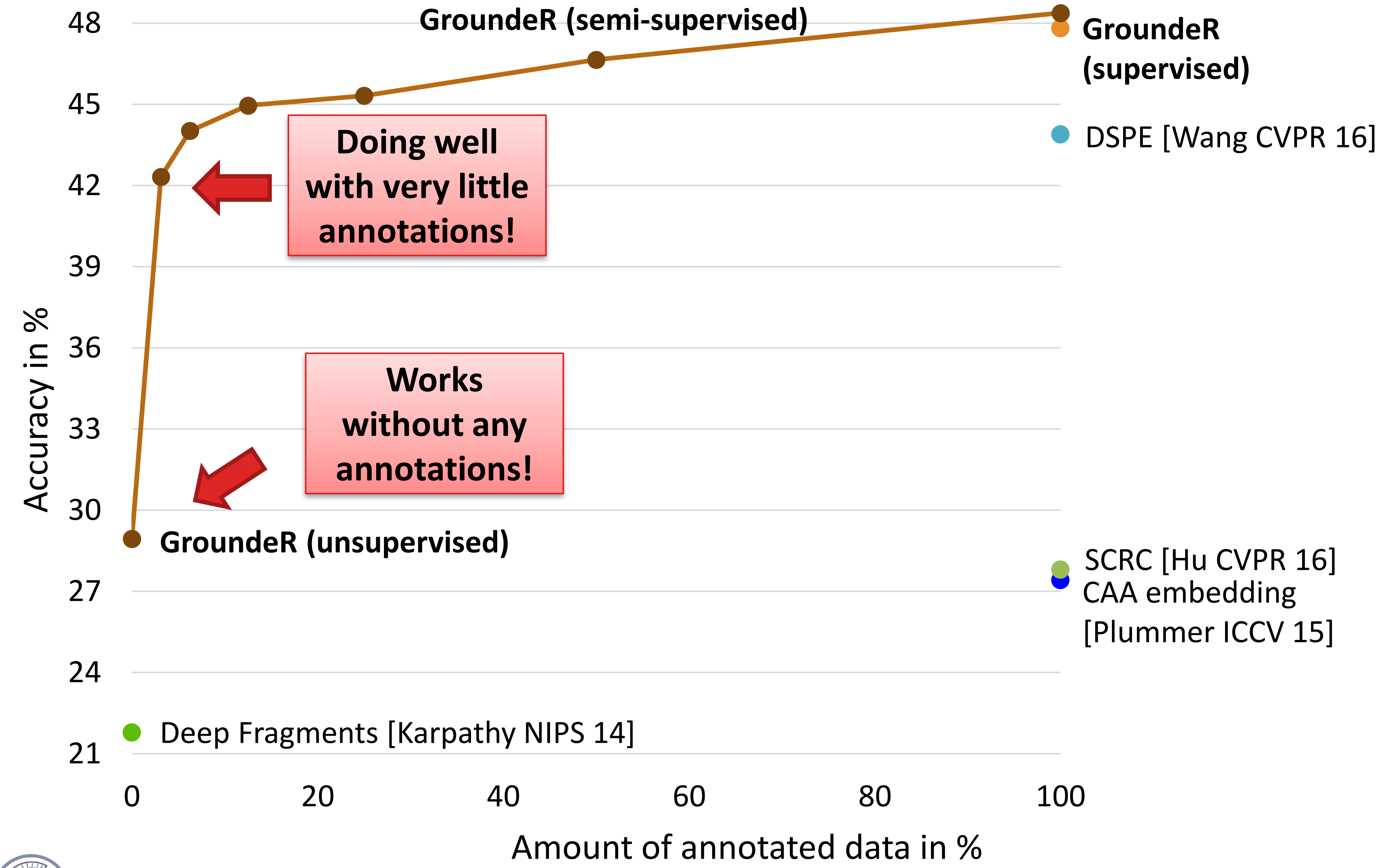


Evaluation: GroundeR on Flickr30k Entities



Works without any annotations!

Evaluation: GroundeR on Flickr30k Entities

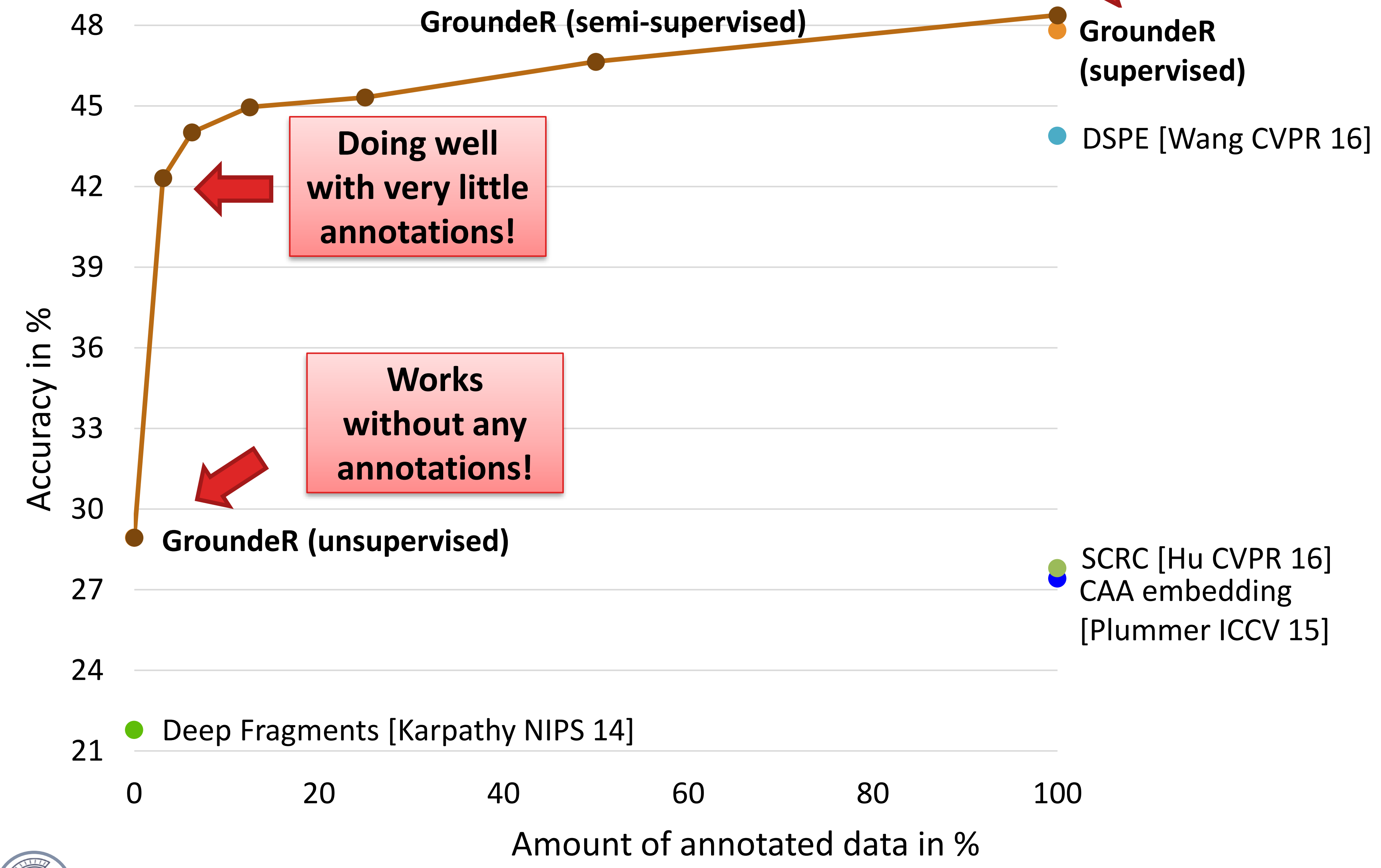


Doing well with very little annotations!

Works without any annotations!

Evaluation: GroundeR on Flickr30k Entities

Semi-supervised better than supervised

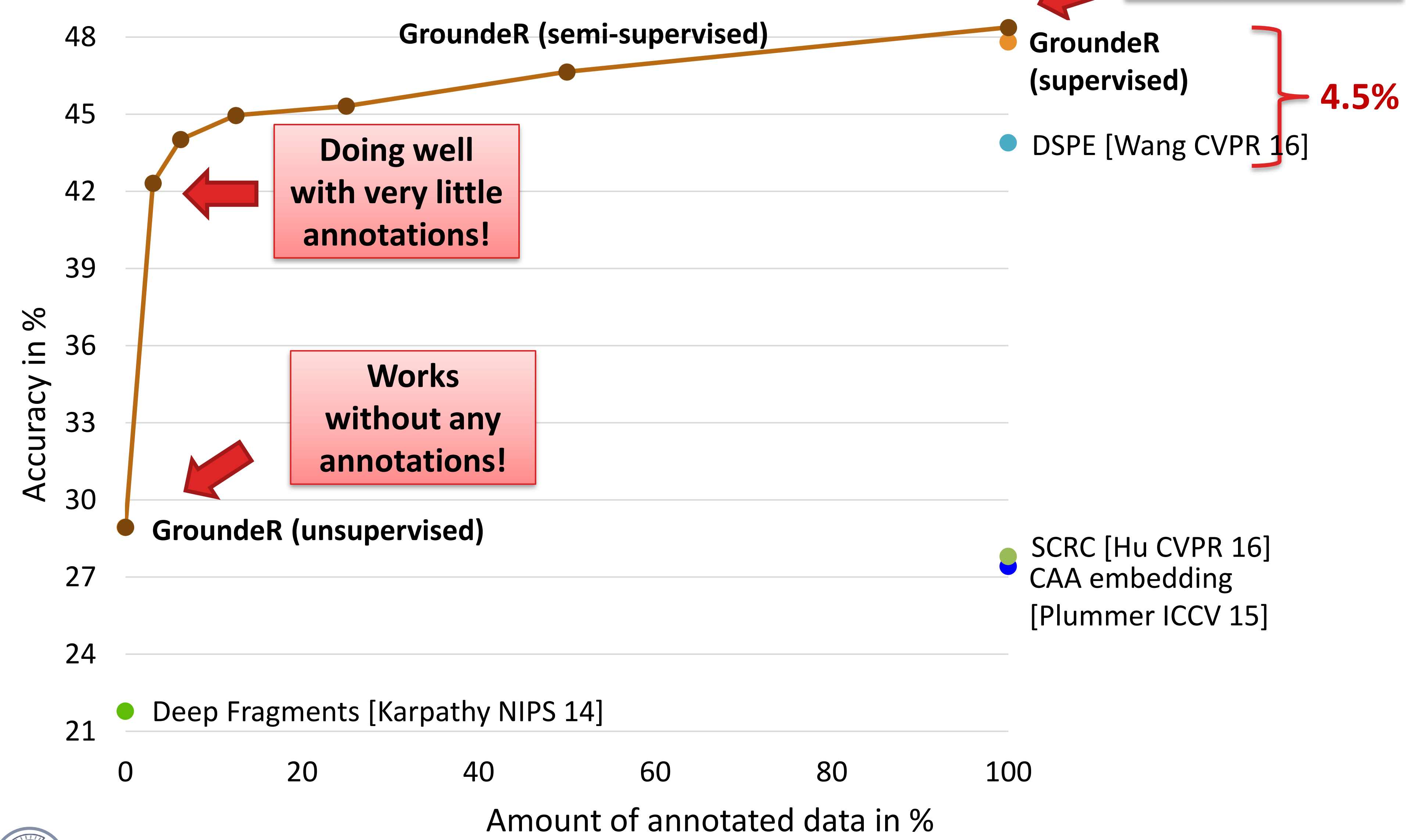


Doing well with very little annotations!

Works without any annotations!

Evaluation: GroundeR on Flickr30k Entities

Semi-supervised better than supervised



Doing well with very little annotations!

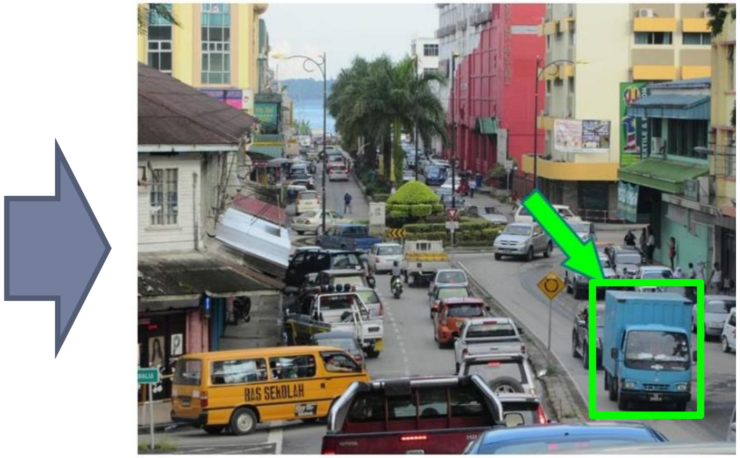
Works without any annotations!

4.5%

Datasets and Experimental setup

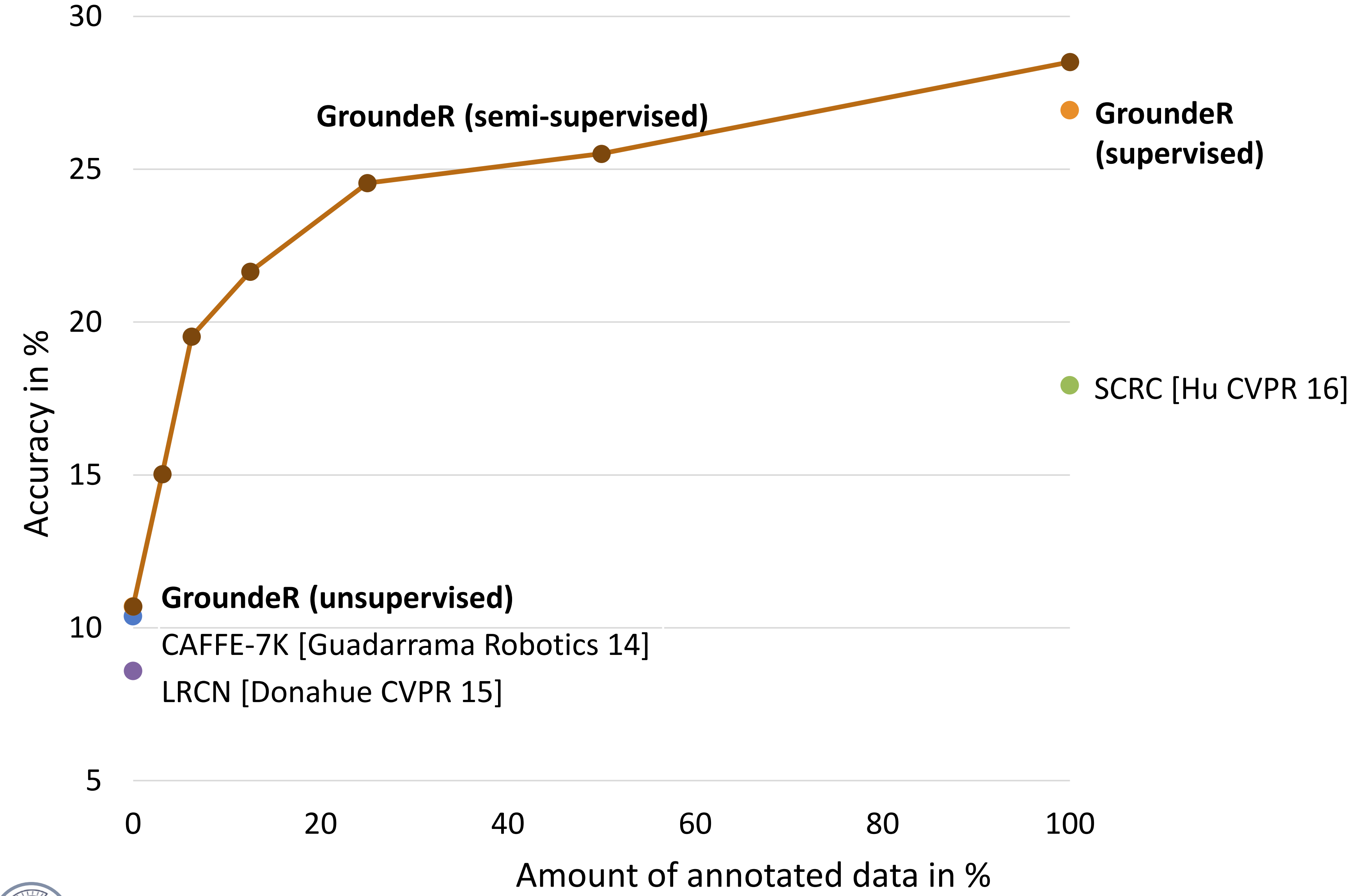
- ReferItGame [Kazemzadeh EMNLP'14]
  - 99K regions & referring expressions

- The blue truck in the bottom right corner
- The light blue truck
- The blue truck on the right

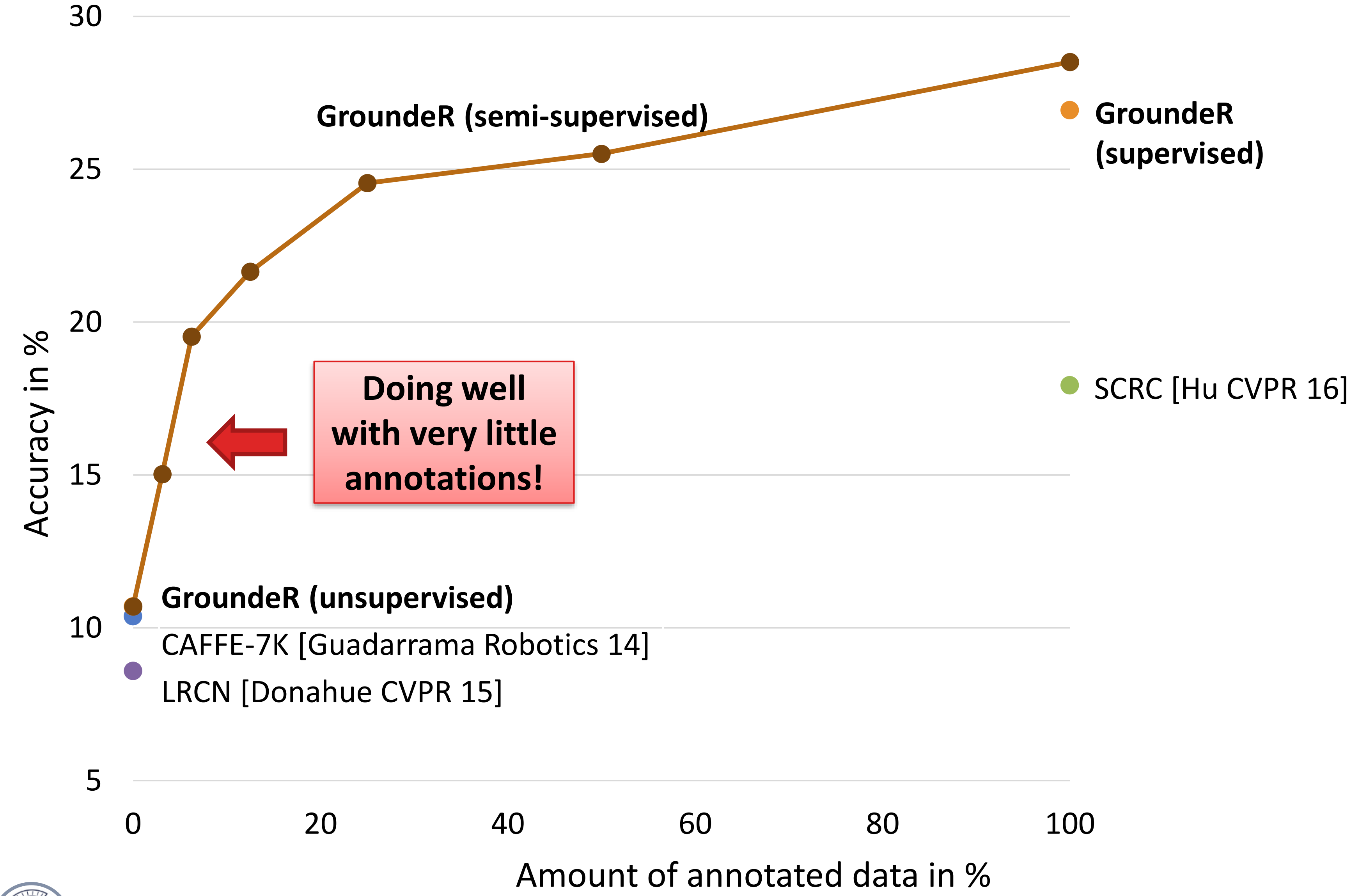


- Experimental setup
  - 100 object proposals
  - VGG16 + spatial feat [Hu CVPR'16]
- Accuracy
  - % phrases: IOU(predicted box, ground-truth box) >= 0.5

Evaluation: GroundeR on ReferItGame

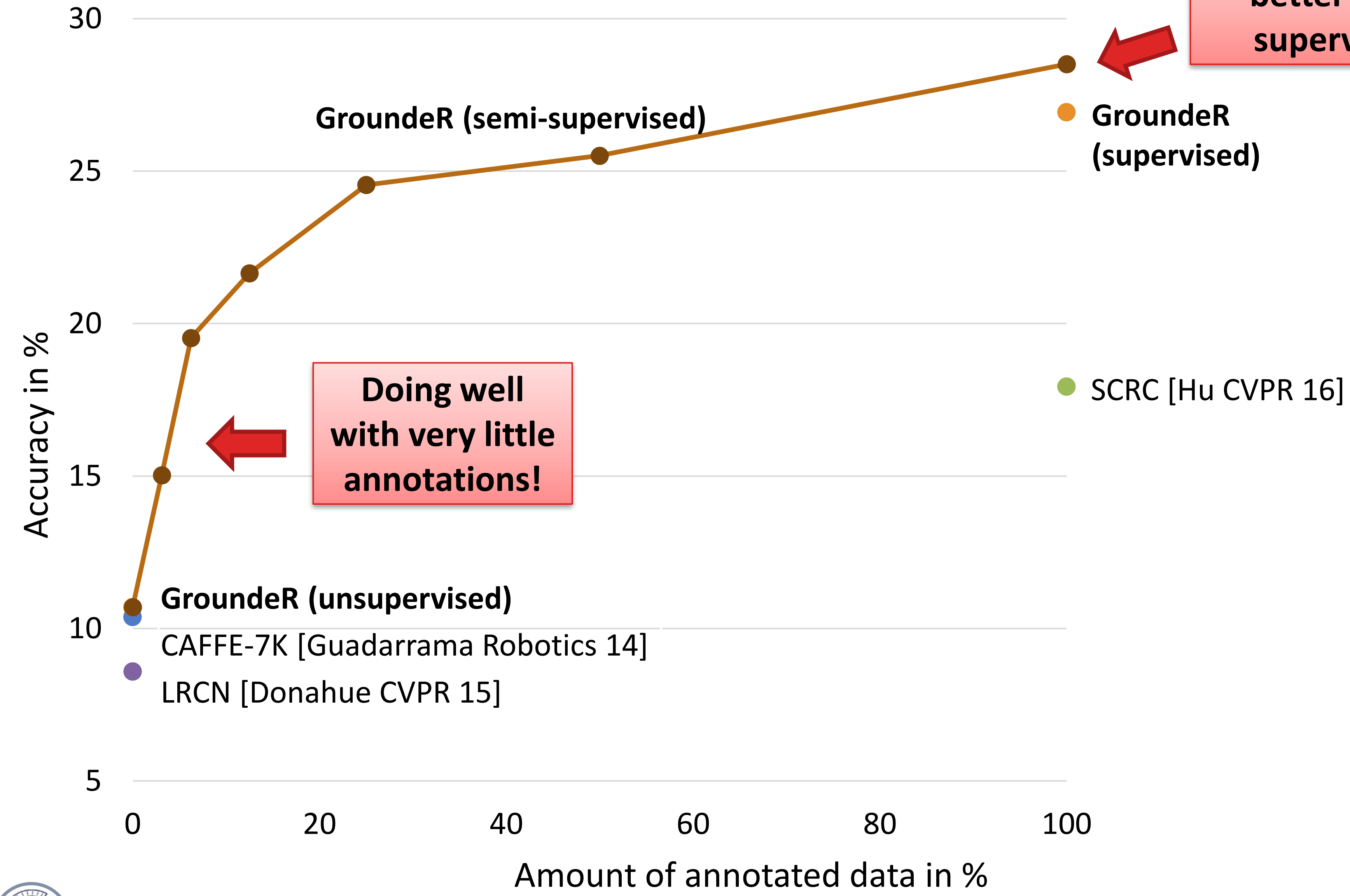


Evaluation: GroundeR on ReferItGame



Doing well with very little annotations!

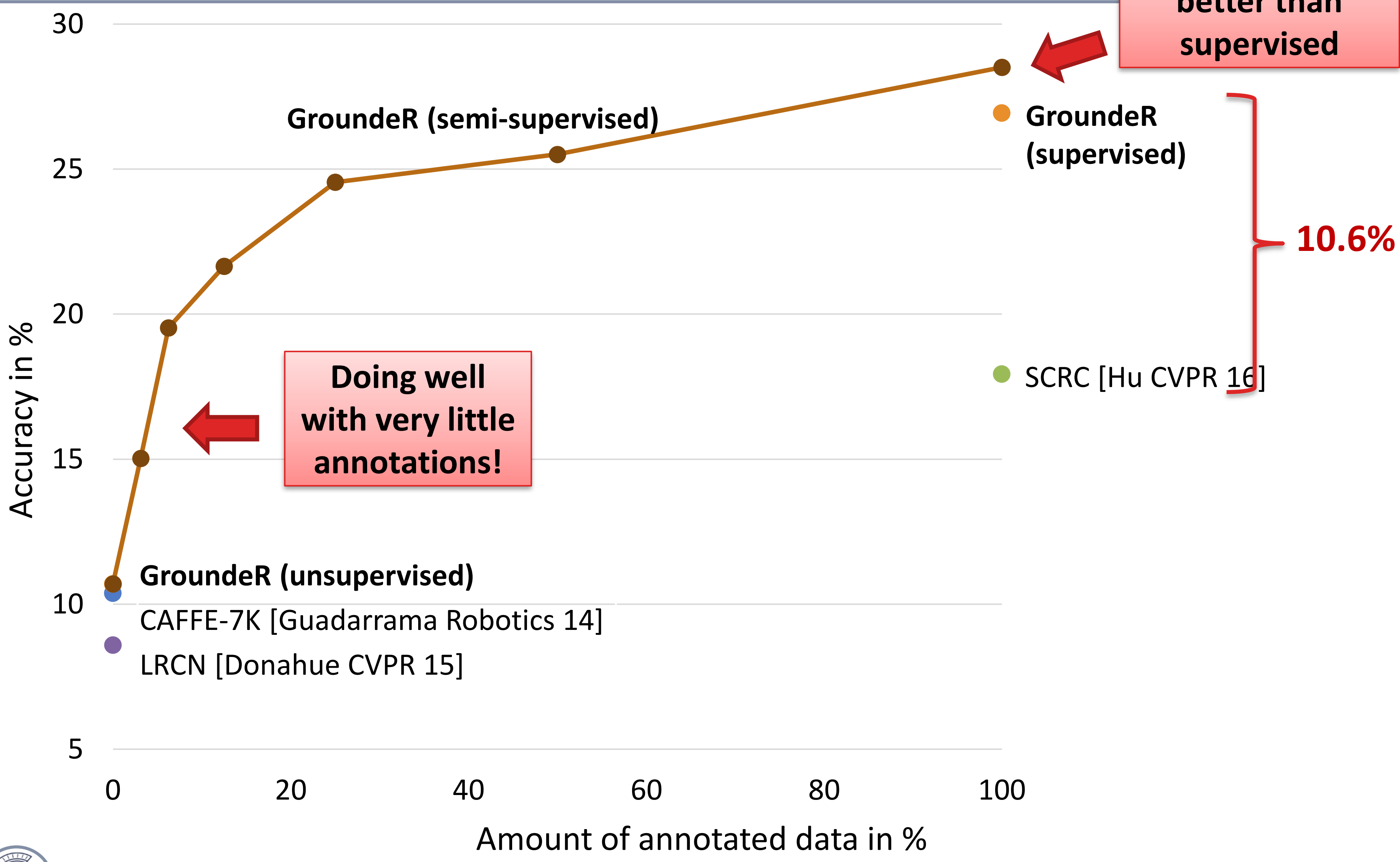
Evaluation: GroundeR on ReferItGame



Semi-supervised better than supervised

Doing well with very little annotations!

Evaluation: GroundeR on ReferItGame



Semi-supervised better than supervised

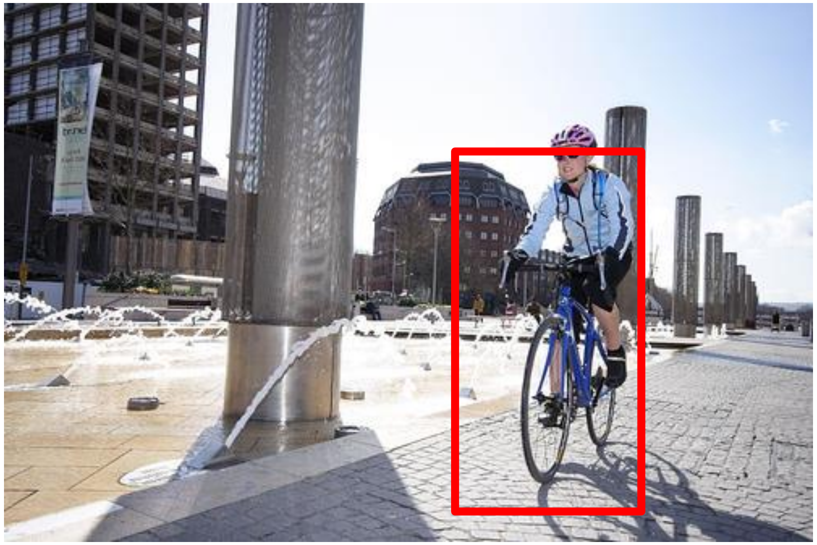
Doing well with very little annotations!

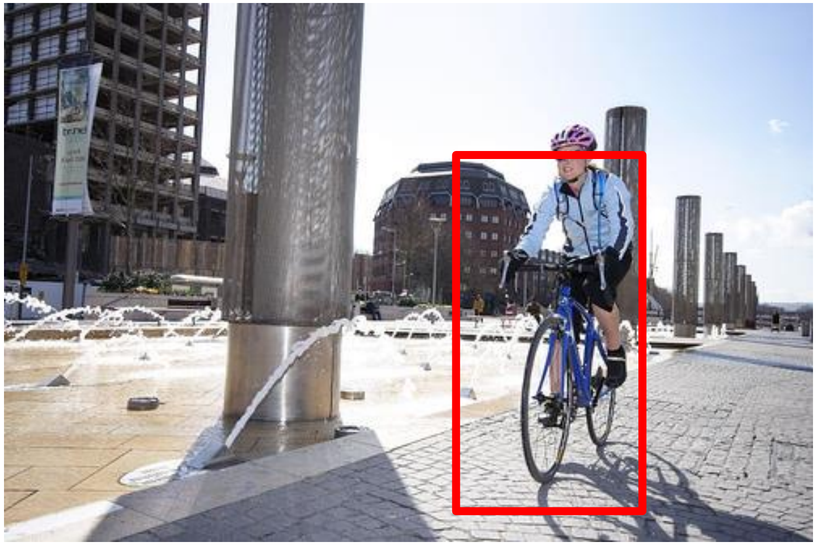
Qualitative results on Flickr30k Entities

GroundeR unsupervised

GroundeR supervised

A woman is riding a bicycle on the pavement.



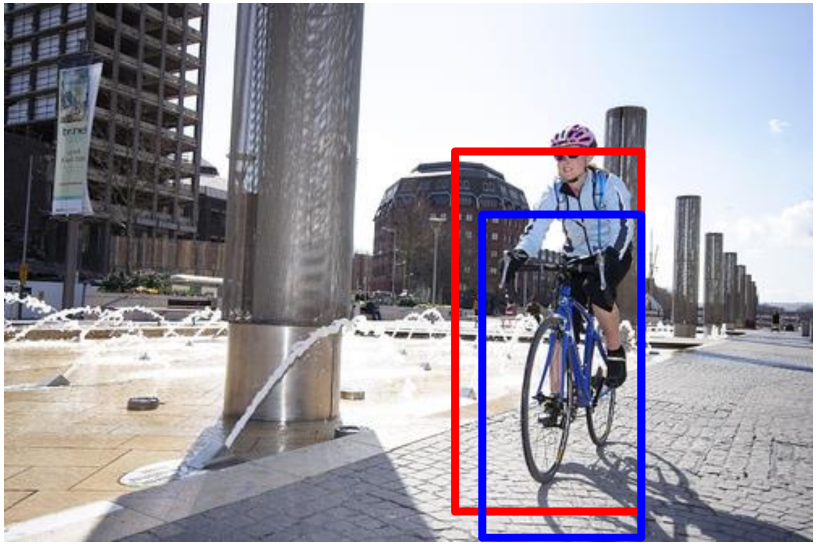


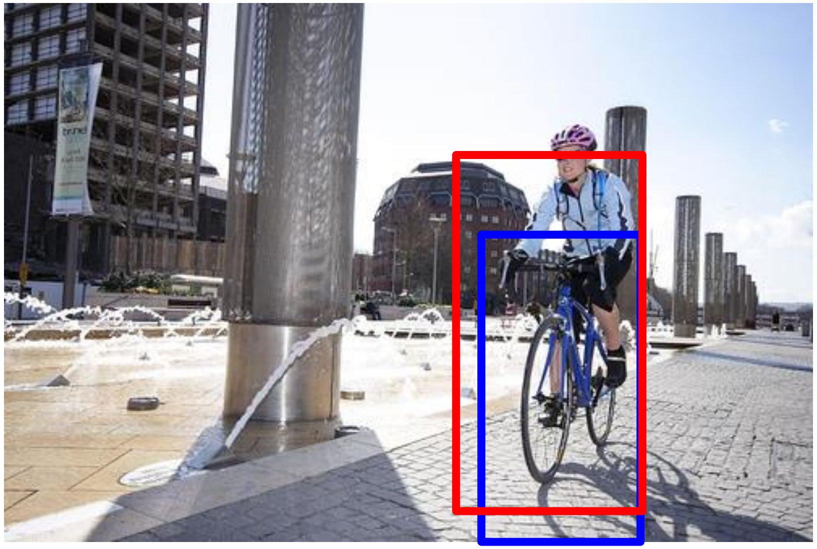
Qualitative results on Flickr30k Entities

GroundeR unsupervised

GroundeR supervised

A woman is riding a bicycle on the pavement.



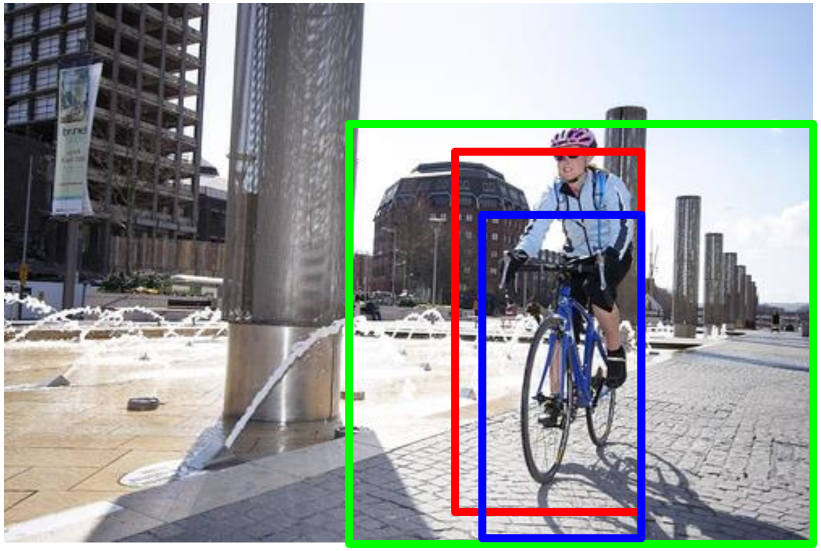


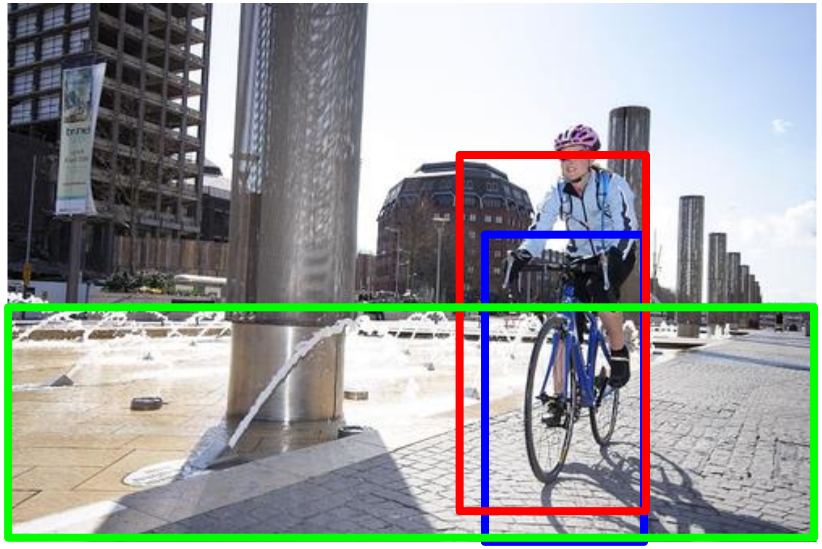
Qualitative results on Flickr30k Entities

GroundeR unsupervised

GroundeR supervised

A woman is riding a bicycle on the pavement.



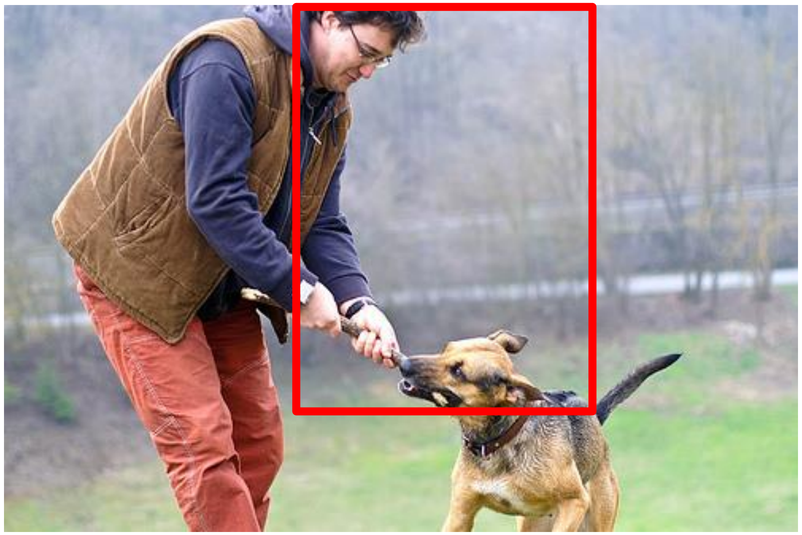


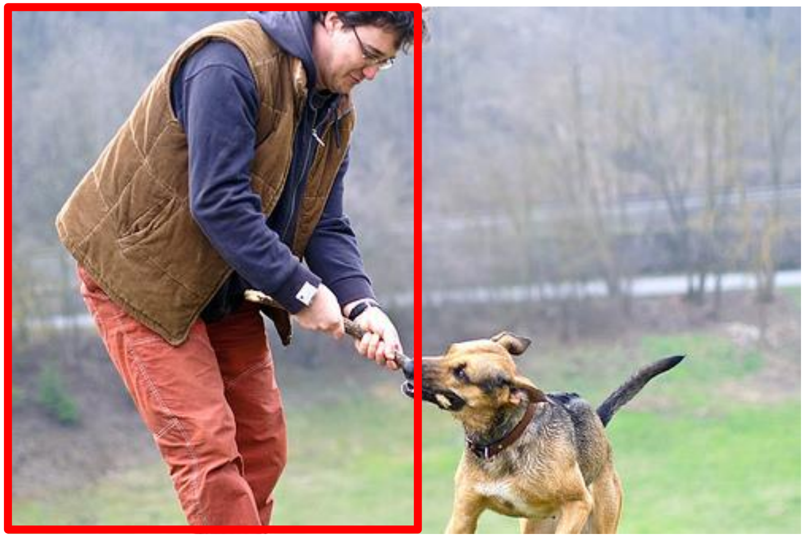
Qualitative results on Flickr30k Entities

SCRC [Hu CVPR 16]

GroundeR semi-supervised (3.12% annot.)

A man in orange pants and brown vest is playing tug-of-war with a dog.



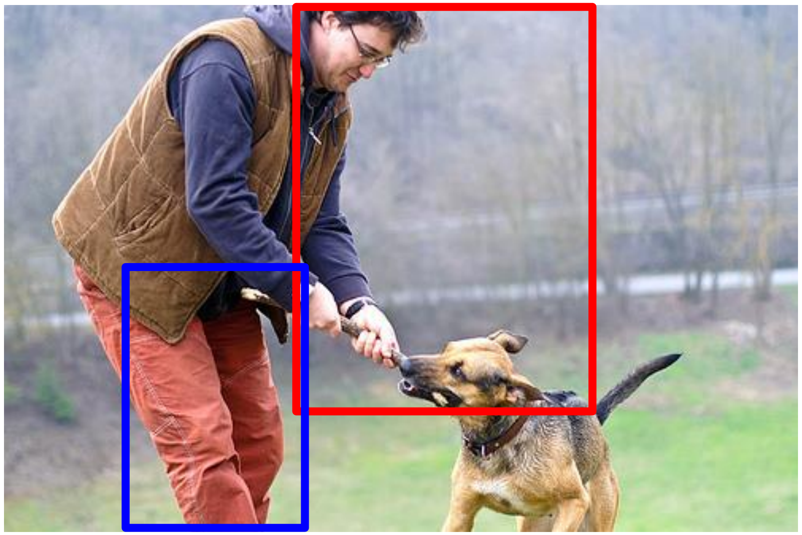


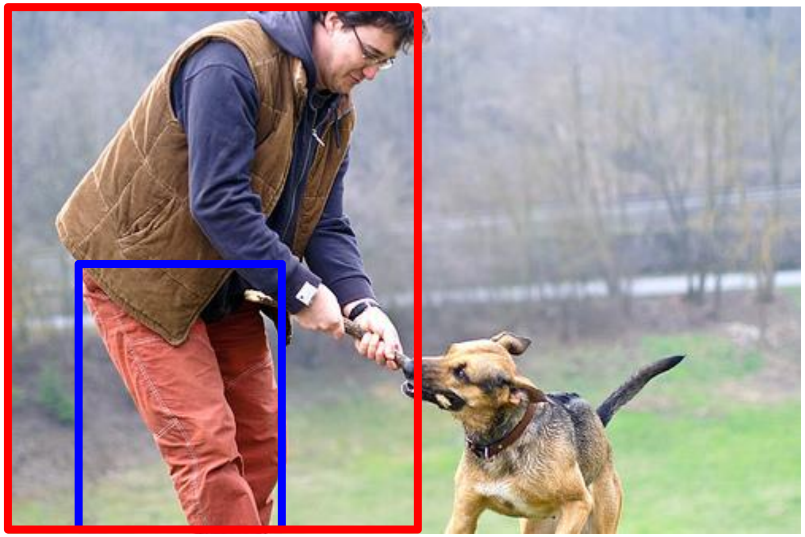
Qualitative results on Flickr30k Entities

SCRC [Hu CVPR 16]

GroundeR semi-supervised (3.12% annot.)

A man in orange pants and brown vest is playing tug-of-war with a dog.



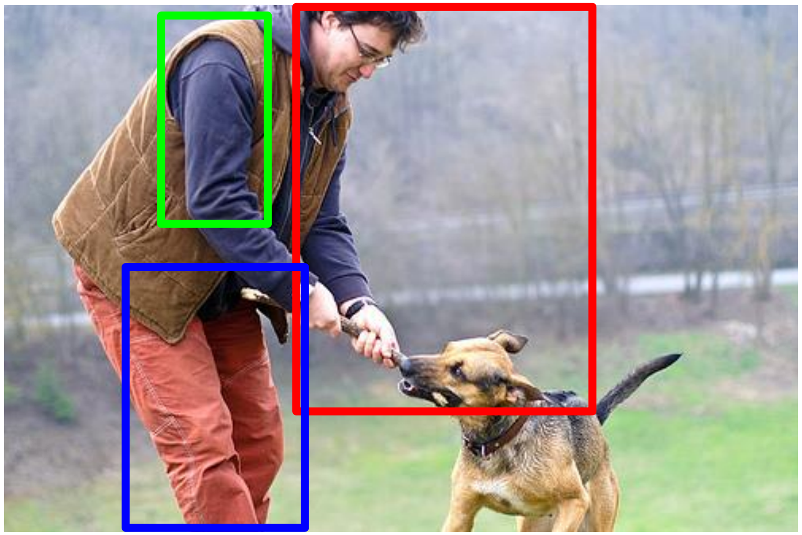


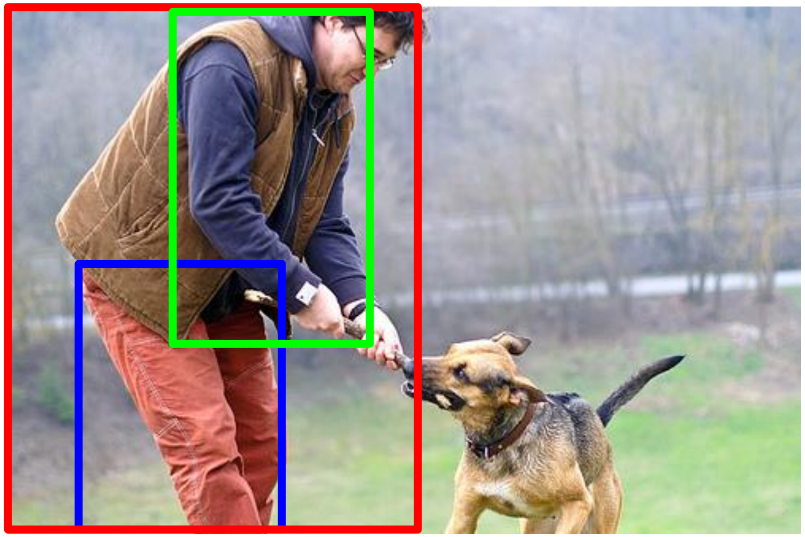
Qualitative results on Flickr30k Entities

SCRC [Hu CVPR 16]

GroundeR semi-supervised
(3.12% annot.)

A man in orange pants and brown vest is playing tug-of-war with a dog.



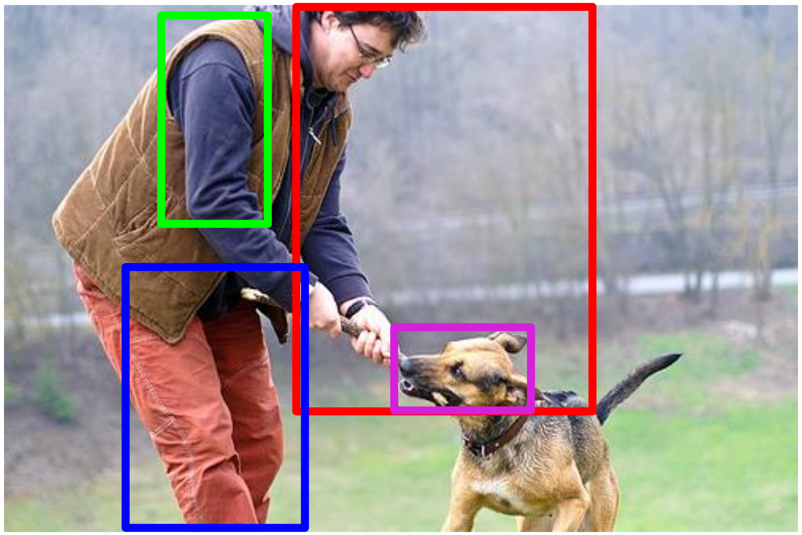


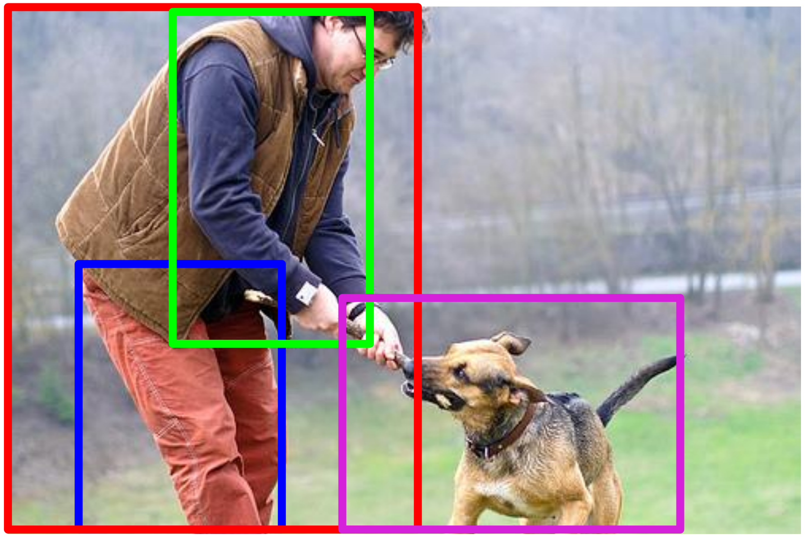
Qualitative results on Flickr30k Entities

SCRC [Hu CVPR 16]

GroundeR semi-supervised
(3.12% annot.)

A man in orange pants and brown vest is playing tug-of-war with a dog.

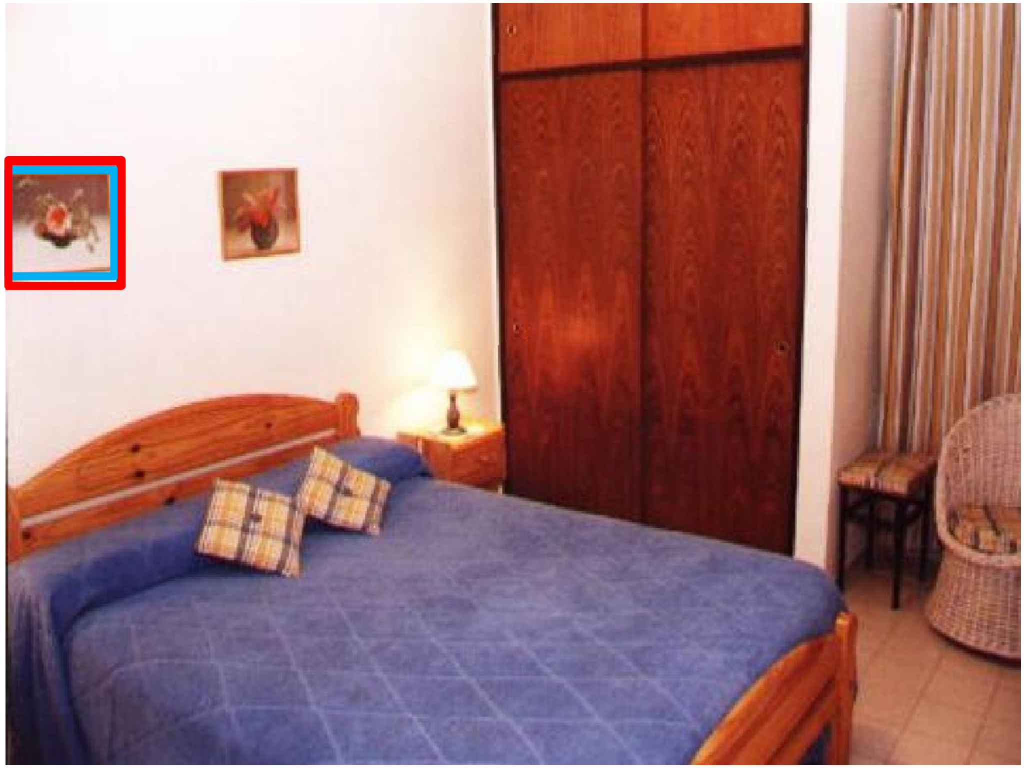




Qualitative results on ReferItGame

GroundeR semi-supervised (12.5% annot.)

picture to the left on the wall

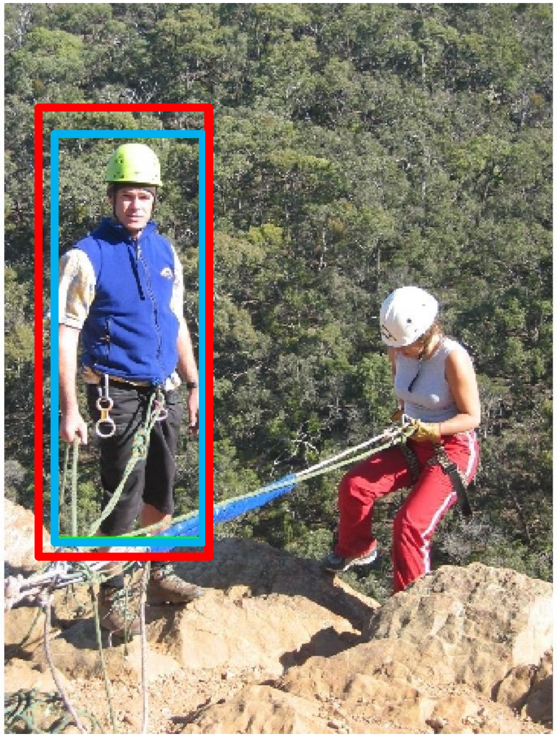


Prediction & ground-truth

Qualitative results on ReferItGame

GroundeR semi-supervised (12.5% annot.)

person in blue

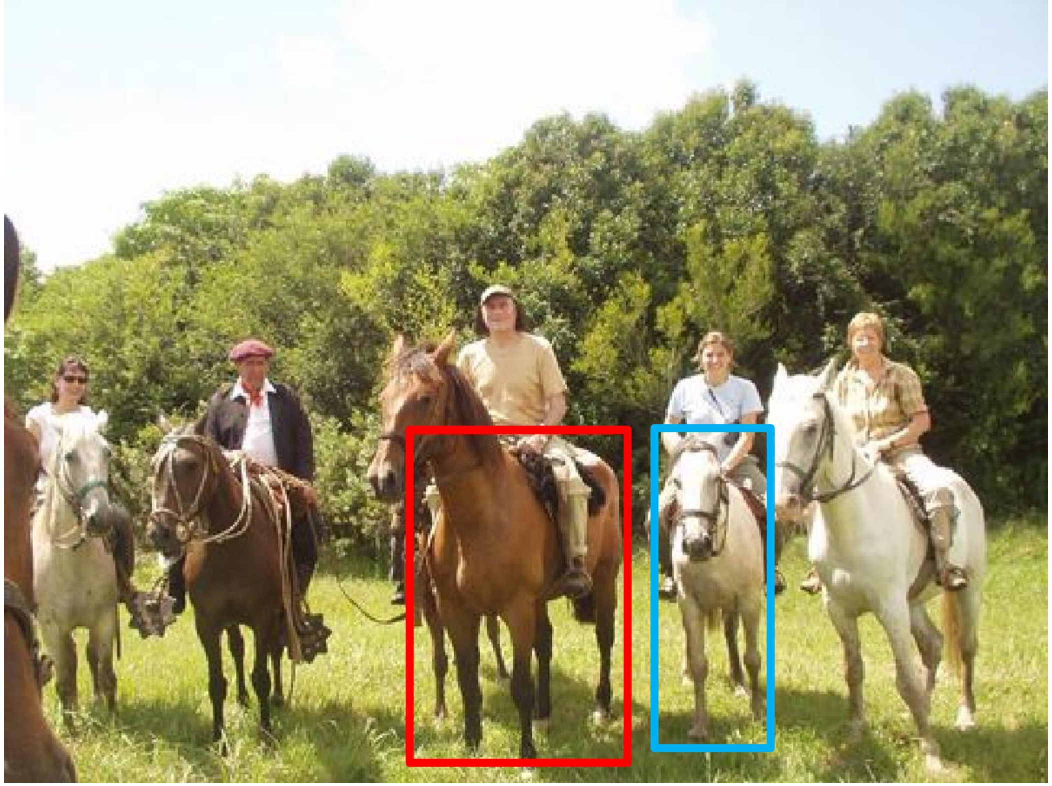


Prediction & ground-truth

Qualitative results on ReferItGame

GroundeR semi-supervised (12.5% annot.)

white horse right of brown horse in middle



Prediction & ground-truth

Conclusions

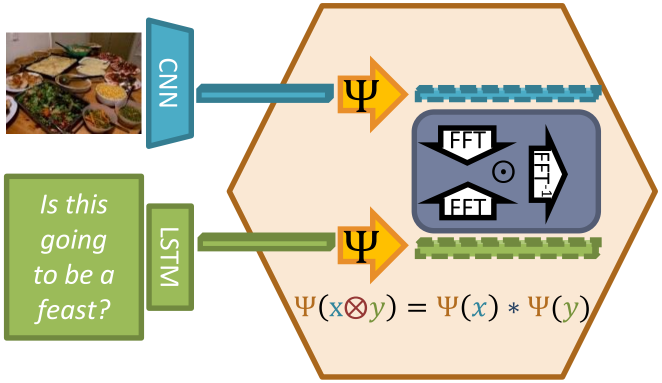
- Unsupervised grounding possible
  - GroundeR with reconstruction objective

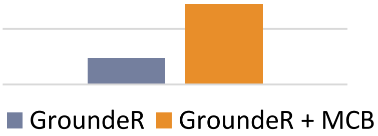
- Semi-supervised GroundeR works best
  - Efficient with little annotations
  - Outperforms fully supervised
  - Outperforms the state-of-the-art

- Possible extensions
  - Jointly reason about multiple phrases
  - Model spatial relations between them

Poster #O-1B-02
Roof Garden

Multimodal Compact Bilinear pooling


CNN
Is this going to be a feast?
LSTM
Ψ
FFT
⊙
FFT
FFT^{-1}
Ψ
$\Psi(x \otimes y) = \Psi(x) * \Psi(y)$

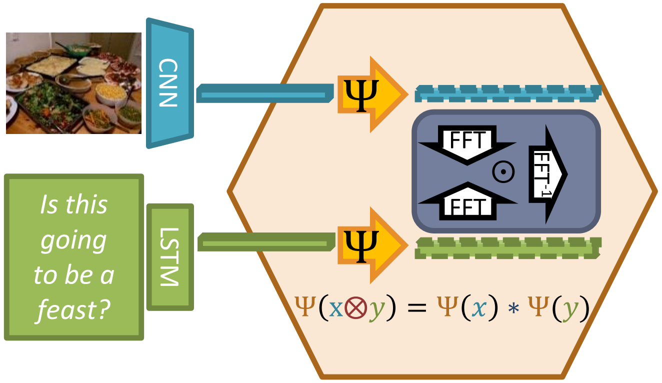


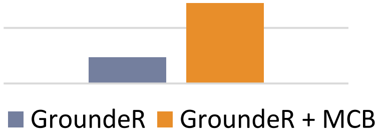
Winner of VQA challenge!

[EMNLP'16]

Poster #O-1B-02
Roof Garden

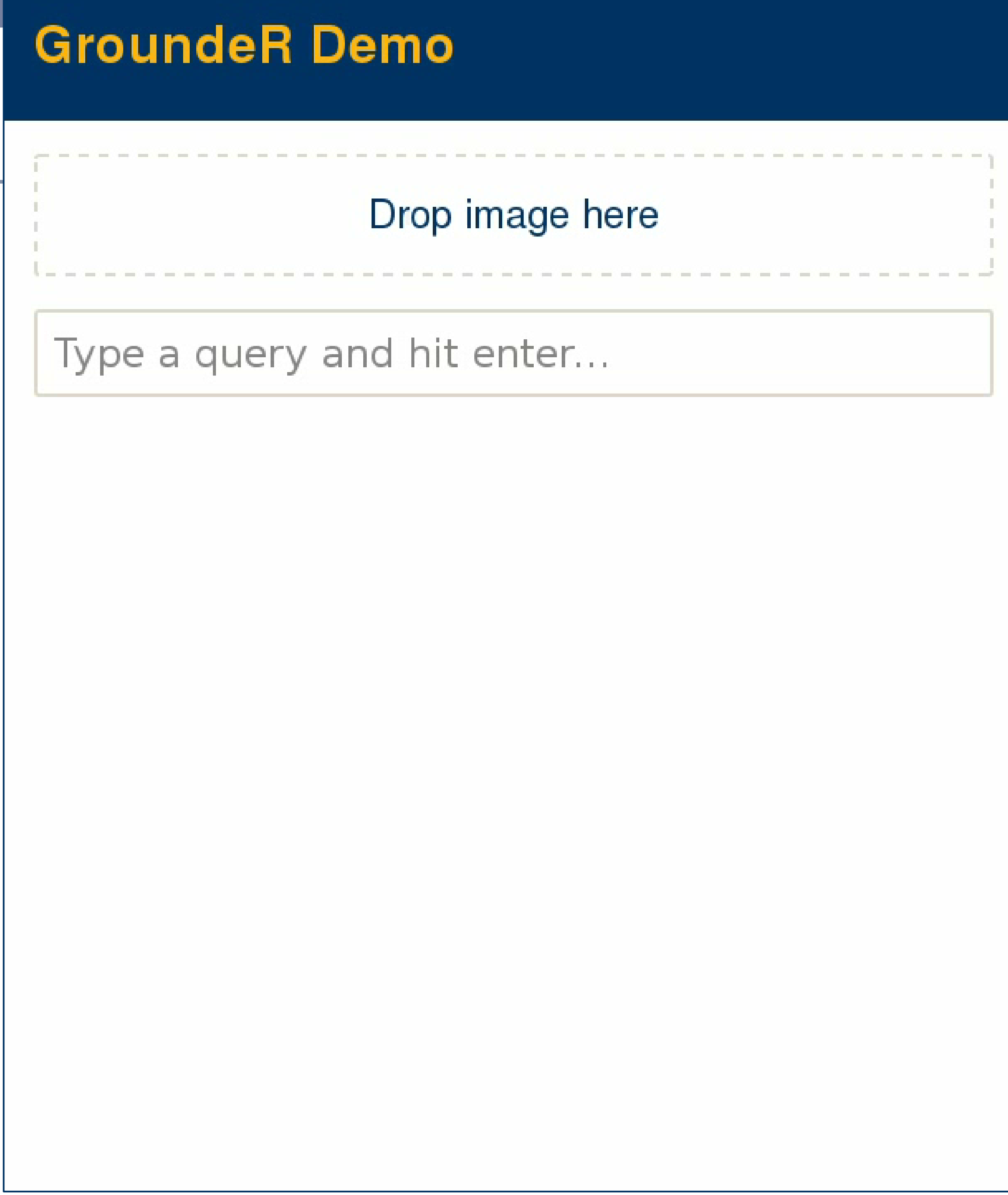
Multimodal Compact Bilinear pooling





Winner of VQA challenge!

[EMNLP'16]



GroundeR Demo

Drop image here

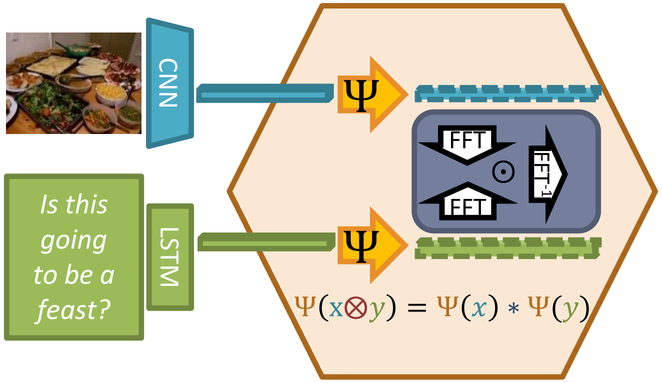
Type a query and hit enter...

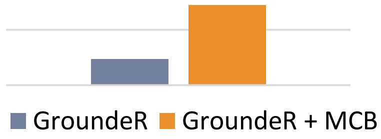
Poster #O-1B-02

Roof Garden

Thank you!

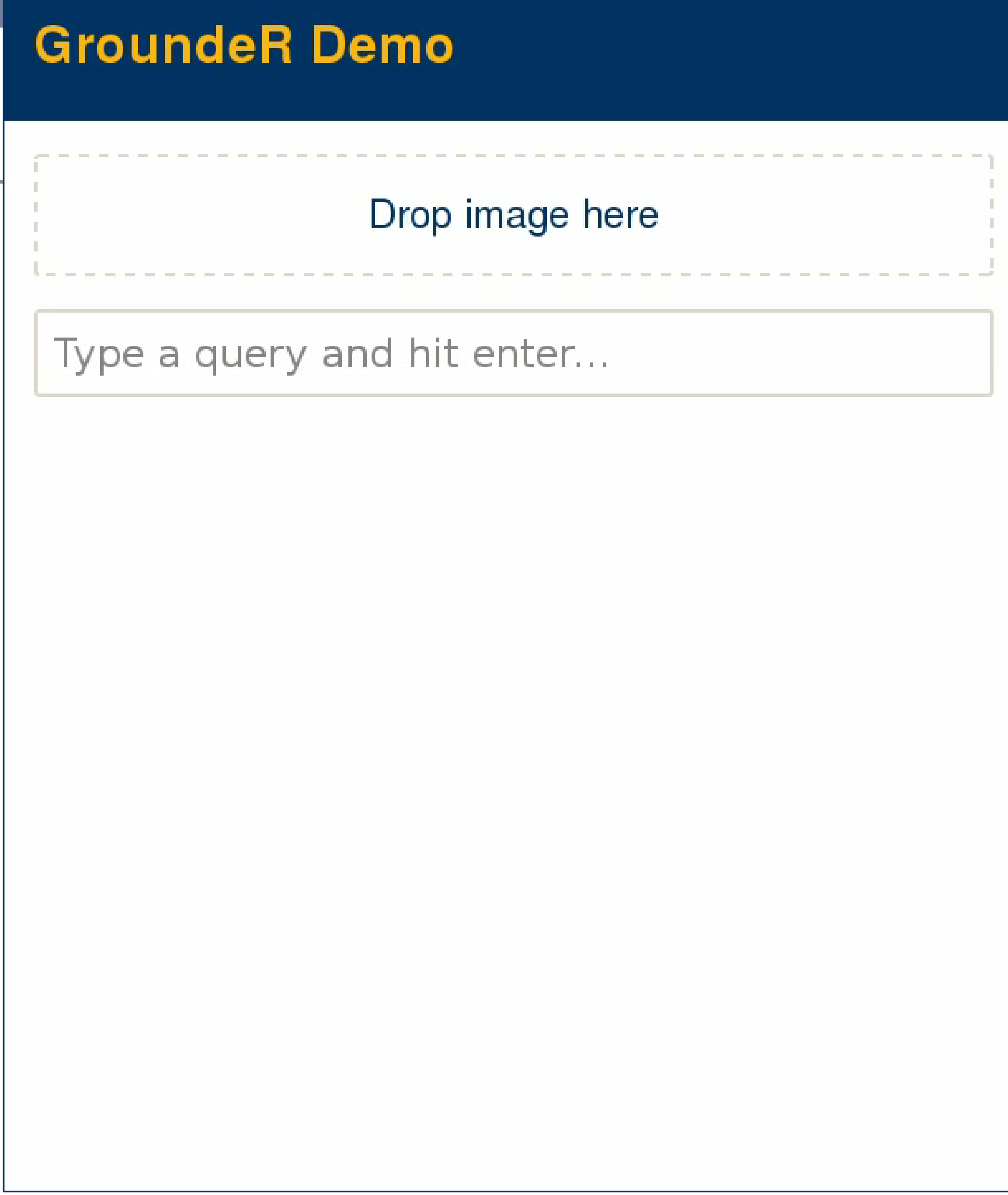
Multimodal Compact Bilinear pooling


CNN
Is this going to be a feast?
LSTM
Ψ
Ψ
FFT
FFT
⊙
FFT⁻¹
$\Psi(x \otimes y) = \Psi(x) * \Psi(y)$



Winner of VQA challenge!

[EMNLP'16]



GroundeR Demo