Relation Networks for Object Detection

Han Hu[1]*, Jiayuan Gu[2]*, Zheng Zhang[1]*, Jifeng Dai[1], and Yichen Wei[1]

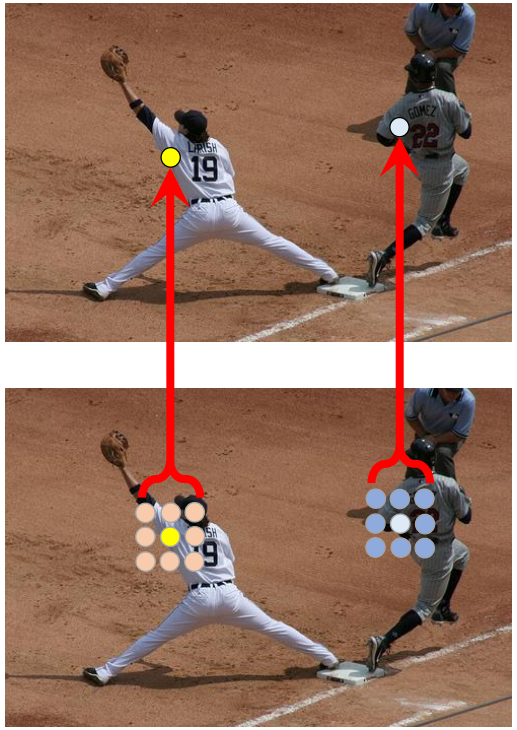
[1]Microsoft Research Asia (MSRA)

[2]Peking University

(*Equal contribution)

Relation Modeling in ConvNets

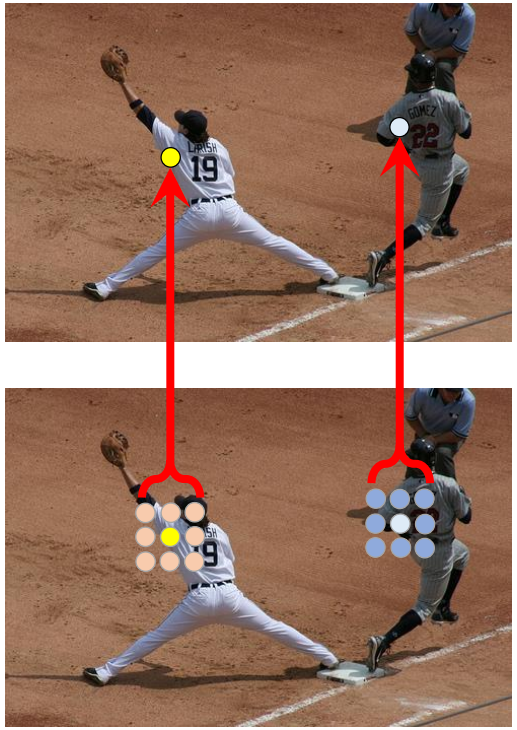
pixel-pixel relation



✓ ***convolution***

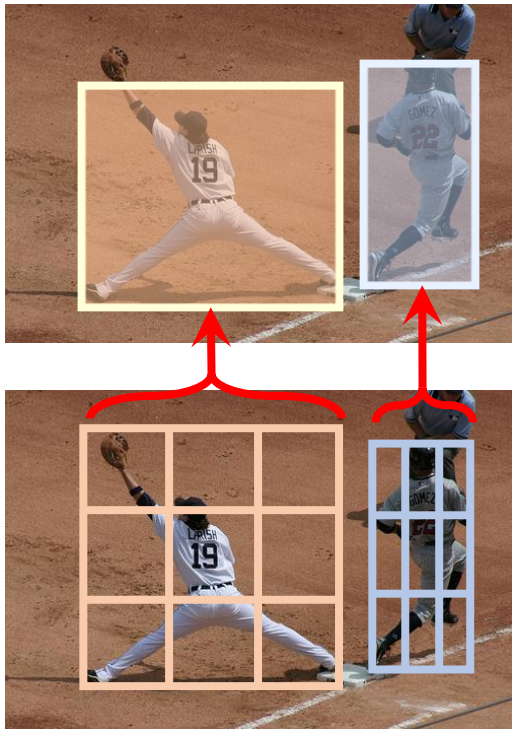
Relation Modeling in ConvNets

pixel-pixel relation



✓ ***convolution***

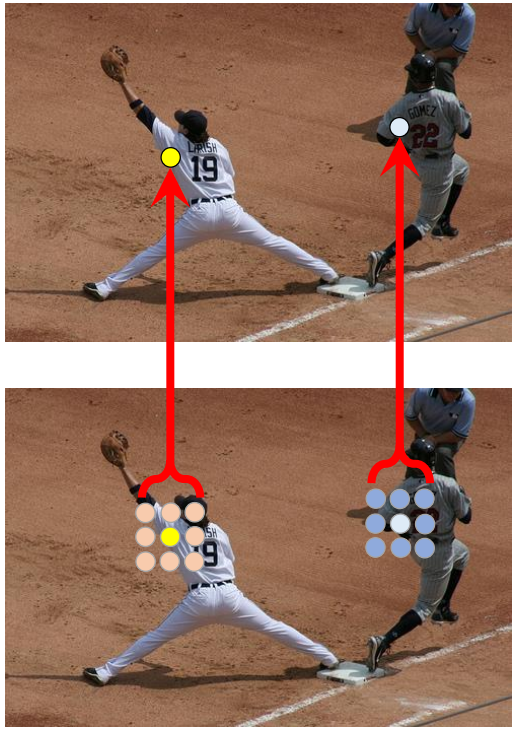
part-part relation



✓ ***RoIPool+FC***

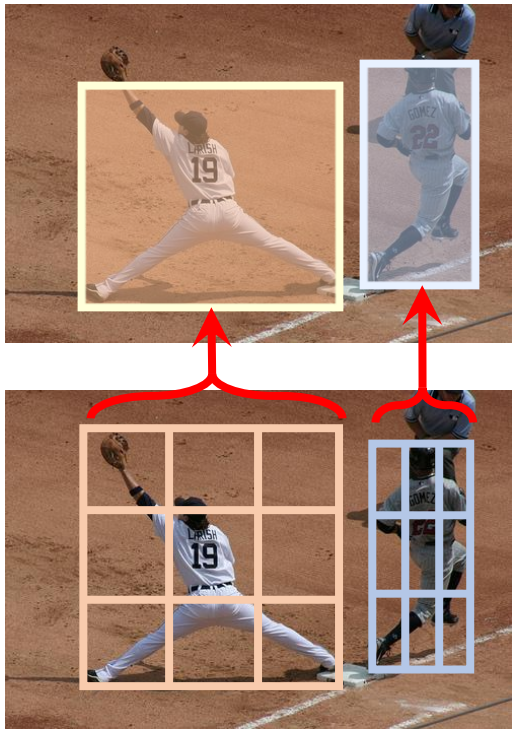
Relation Modeling in ConvNets

pixel-pixel relation



✓ ***convolution***

part-part relation



✓ ***RoIPool+FC***

Effective and Easy to use

✓Parallel

✓Learnable

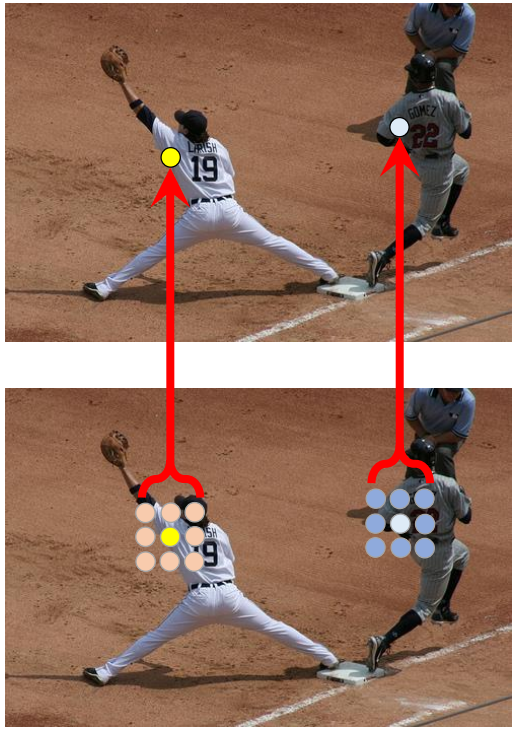
✓Require no relation supervision

✓Translational invariant

✓Stackable (convolution)

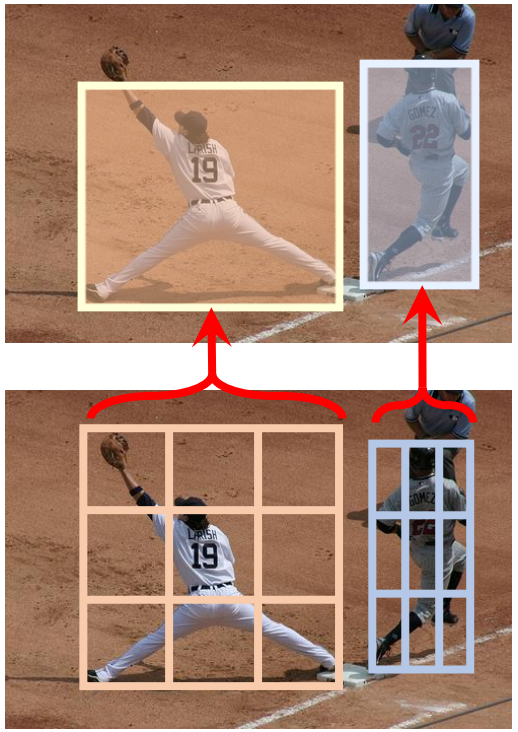
Relation Modeling in ConvNets

pixel-pixel relation



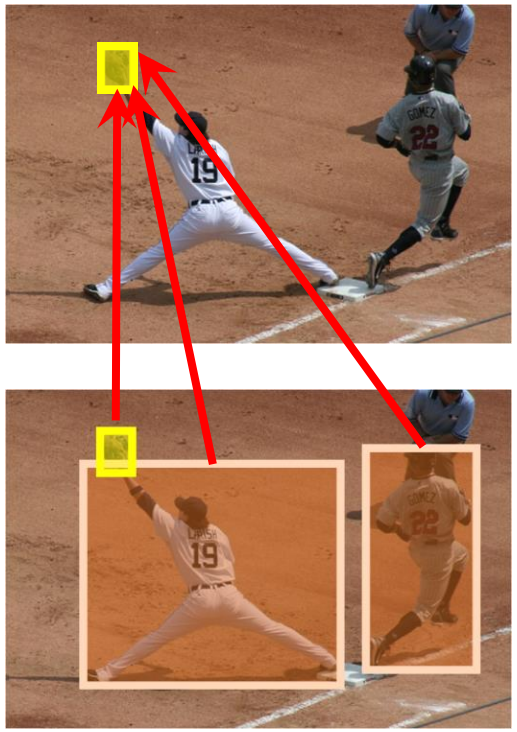
✓ *convolution*

part-part relation



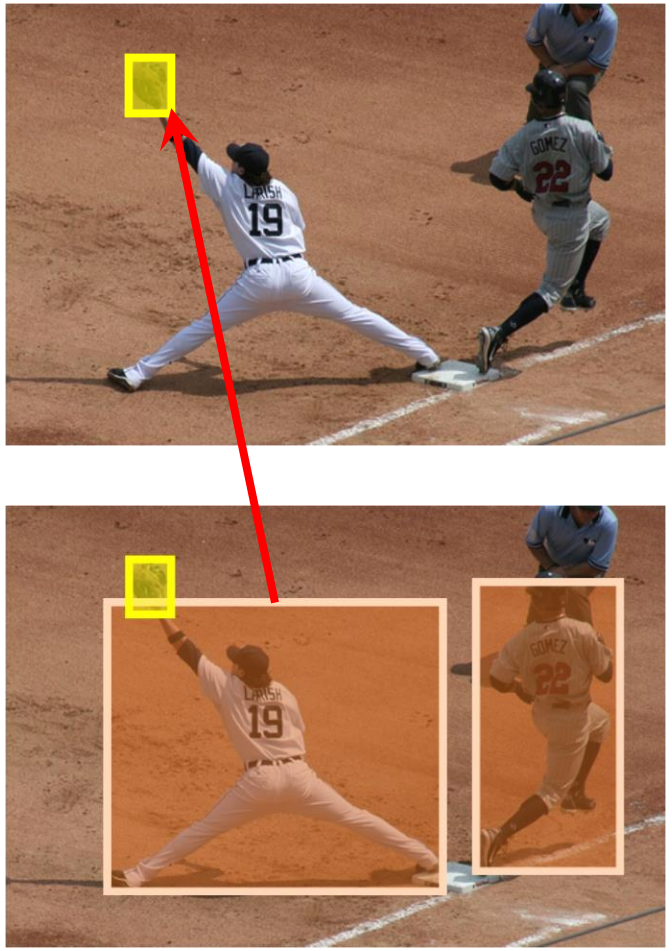
✓ *RoIPool+FC*

object-object **relation**



?

Well Recognized Problem

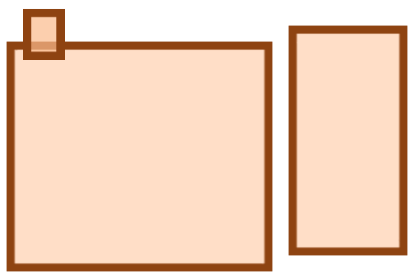


It is much easier to detect the ***glove*** if we know there is a ***person***.

Rarely Studied in Deep Learning Era

pixels

region parts

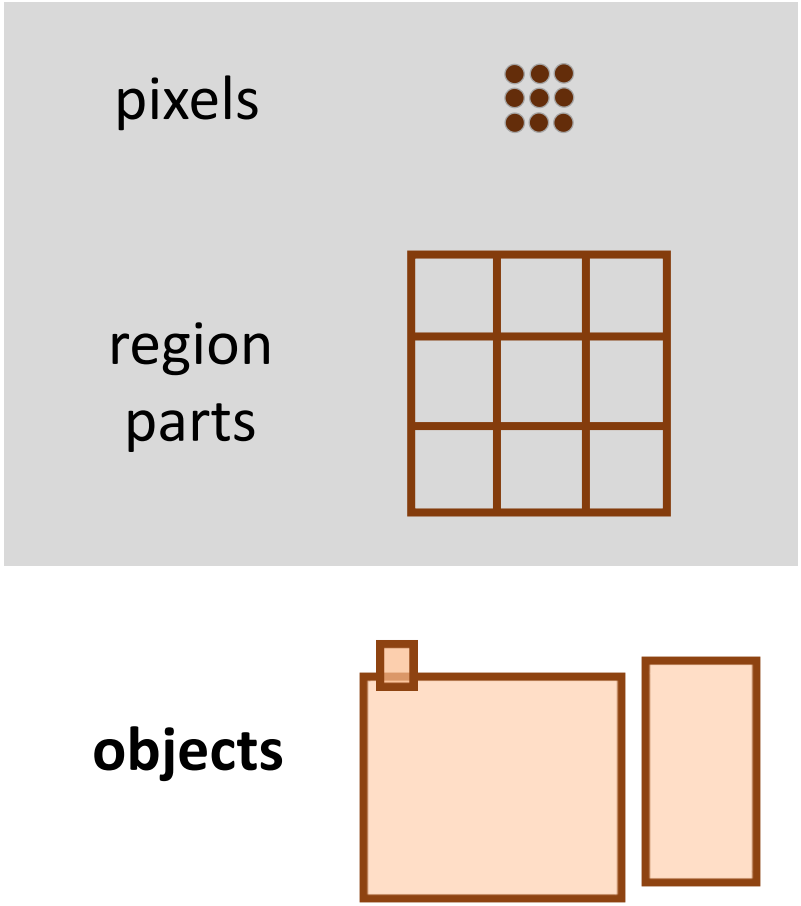


objects

Irregularities of objects

- At arbitrary image locations
- Of different scales
- Within different categories
- Of varying number across different images

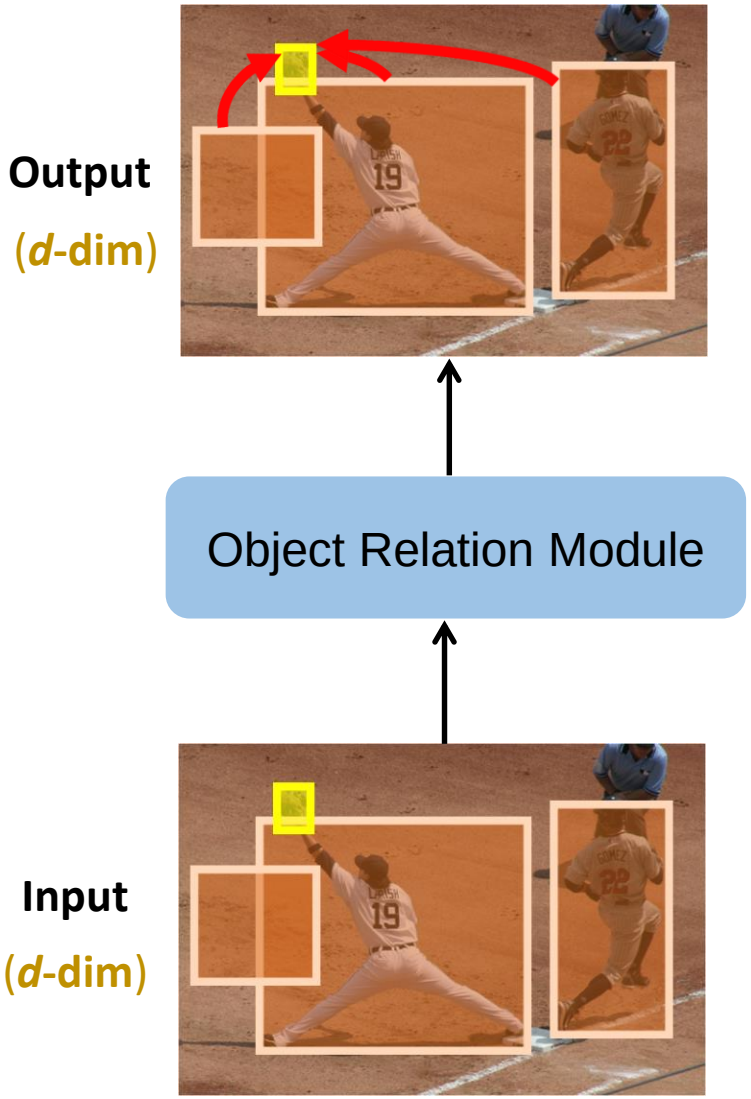
Rarely Studied in Deep Learning Era



Irregularities of objects

- At arbitrary image locations
- Of different scales
- Within different categories
- Of varying number across different images

Goal



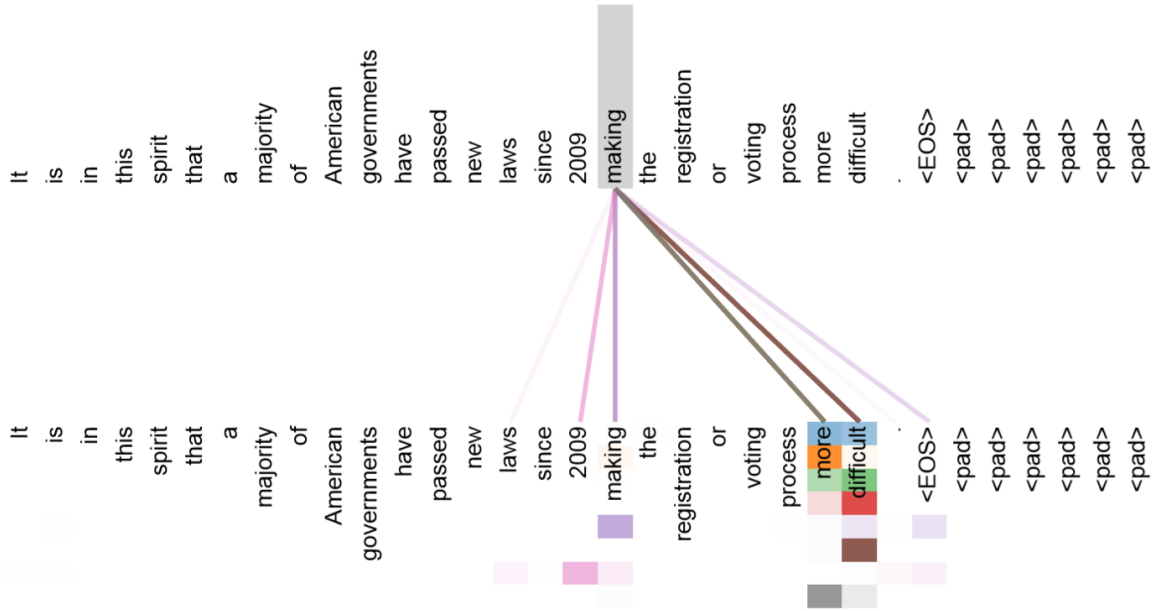
Goal: design a simple module to model object-object relation

Effective and Easy to use

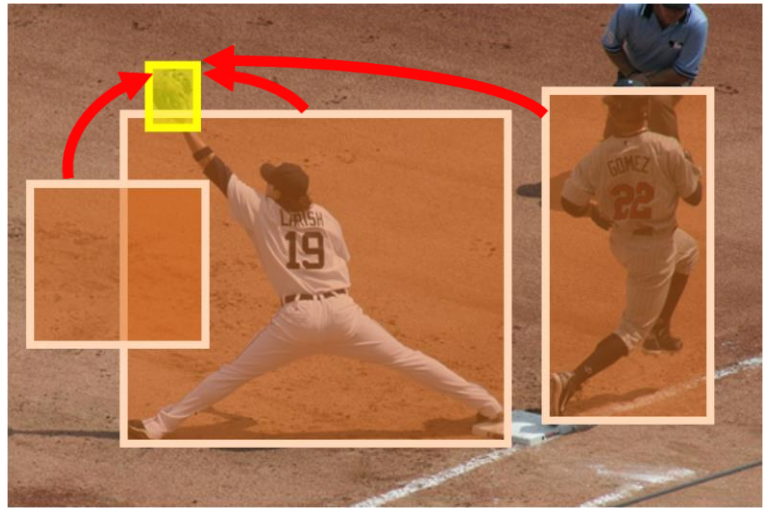
- ✓Parallel
- ✓Learnable
- ✓Require no relation supervision
- ✓Translational invariant
- ✓In-place, stackable

Object Relation Module

- **Extension** of *attention module*



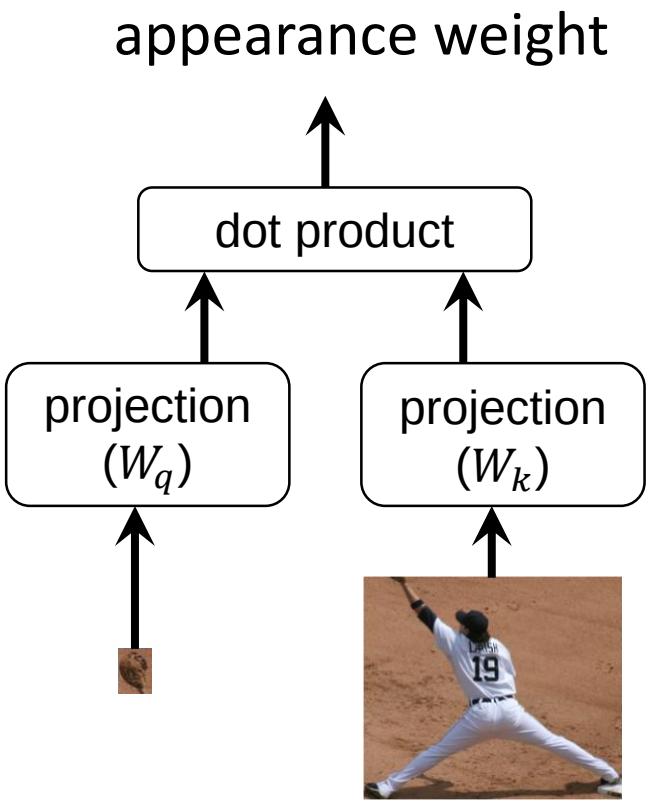
word-word relation
(**1D sequential**)



object-object relation
(**2D irregular**)

Left figure credit by A. Vaswani et al.

Relation between Two Objects

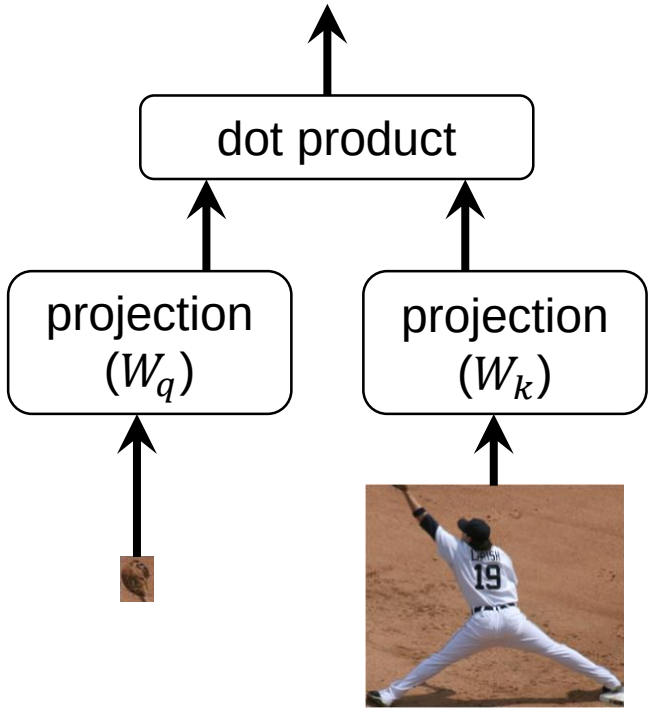


in standard ***attention*** module

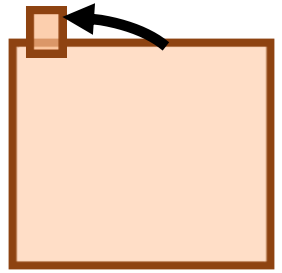
Relation between Two Objects

- A **novel** **geometric** weight

appearance weight



app. **+** **geometric** weight



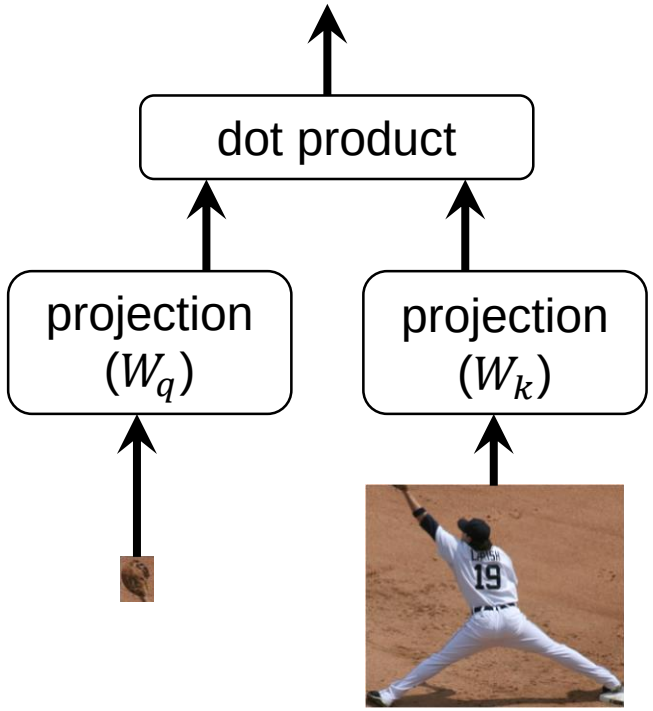
in standard ***attention*** module

in **object relation** module

Relation between Two Objects

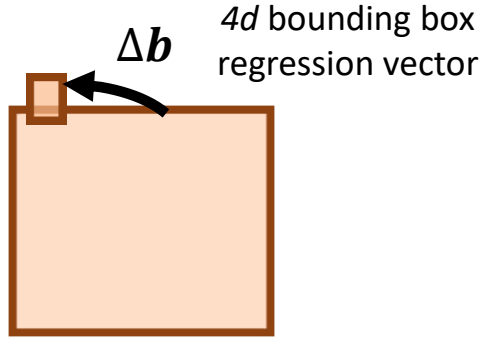
- A **novel** **geometric** weight

appearance weight



in standard ***attention*** module

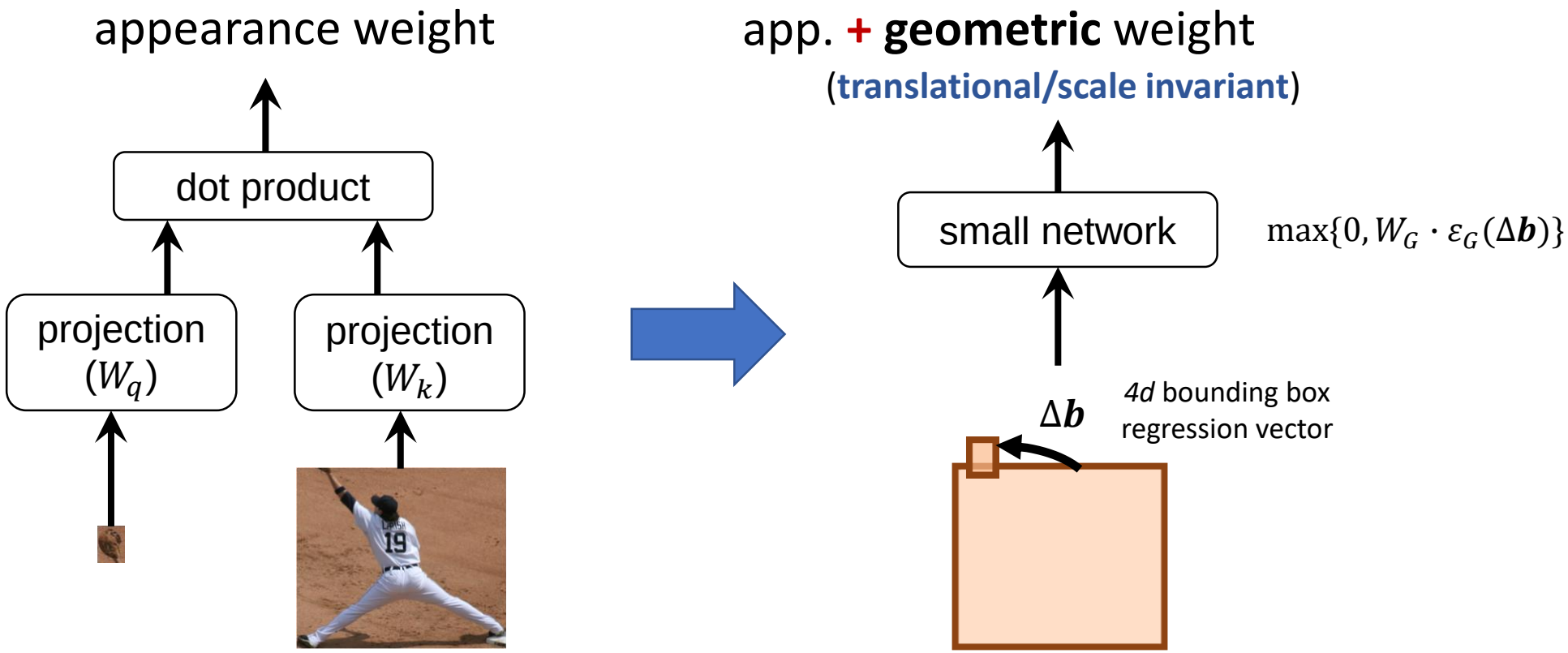
app. **+** **geometric** weight



in **object relation** module

Relation between Two Objects

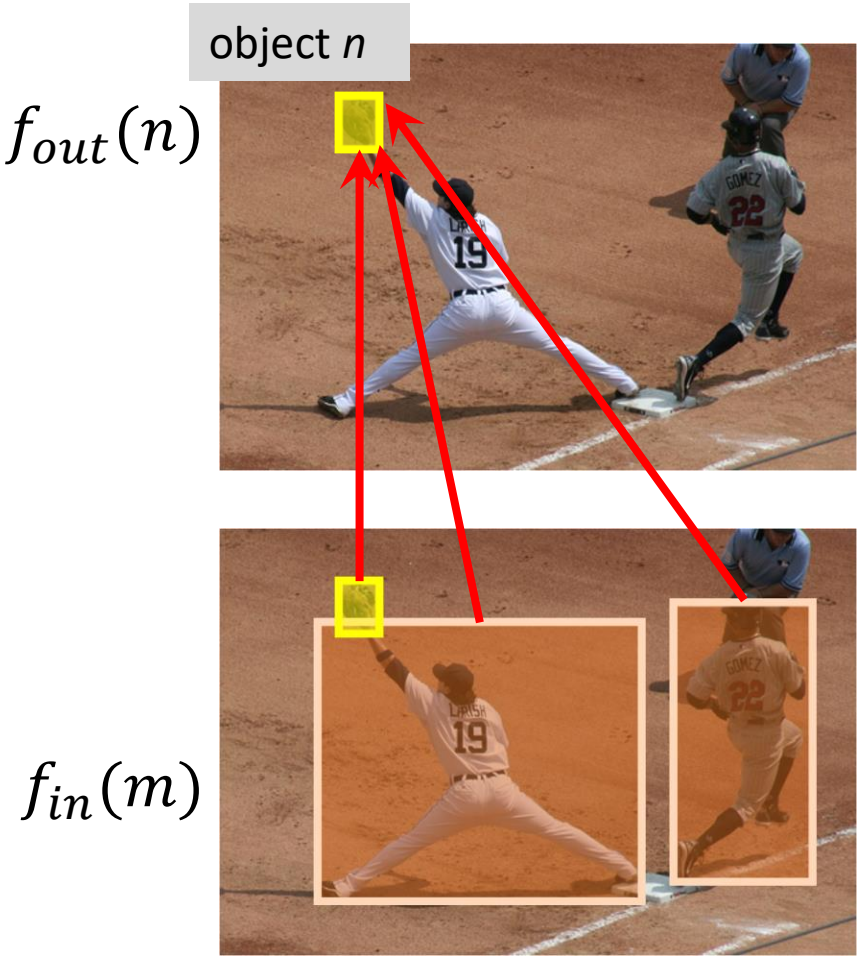
- A **novel** **geometric** weight

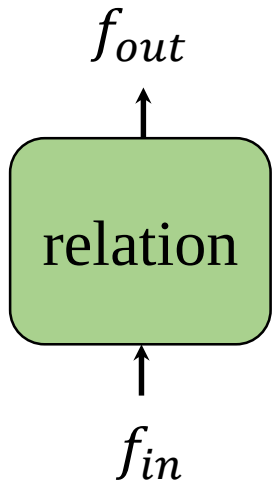


in standard ***attention*** module

in **object relation** module

Relation Aggregation

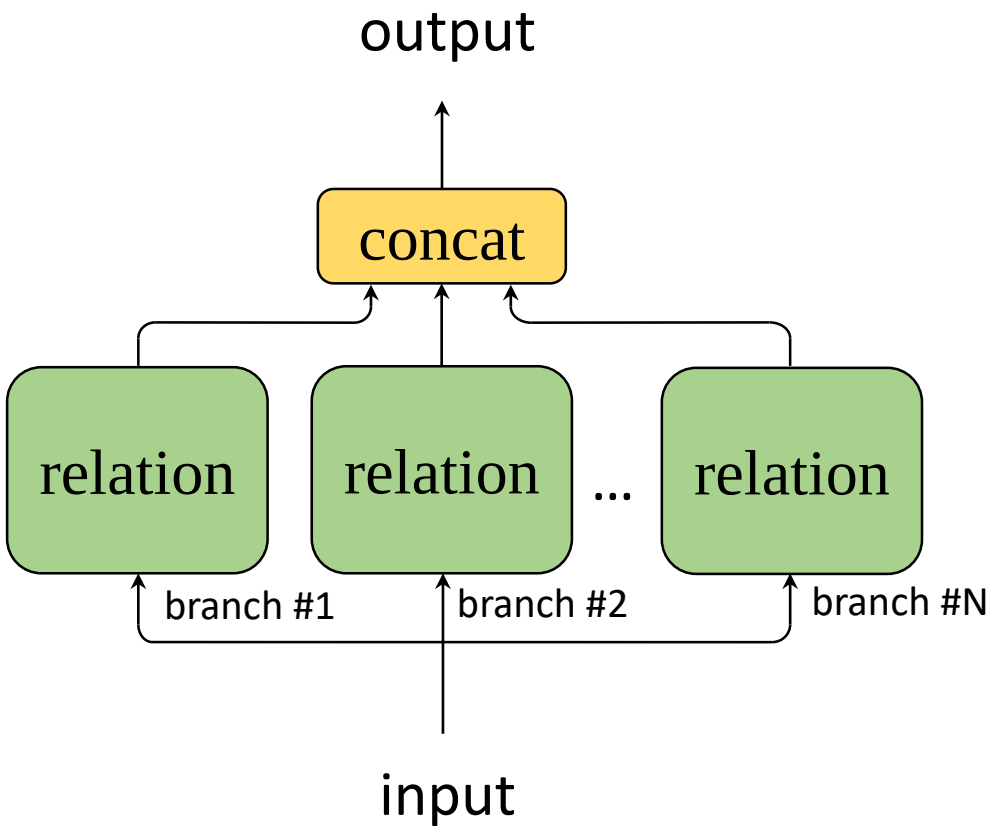




appearance + geometric weight

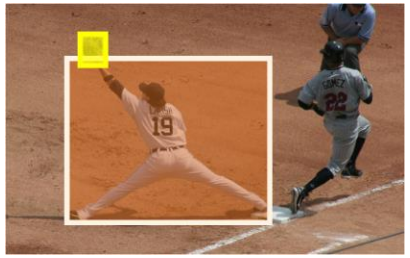
$$f_{out}(n) = \sum_{m} \omega(m,n) \cdot f_{in}(m)$$

Multi-Branch Relation

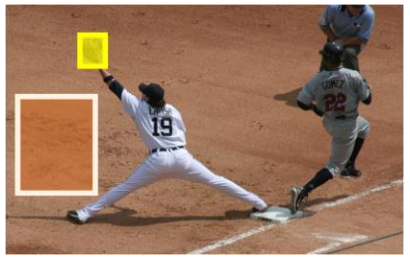


Multi-Branch Relation

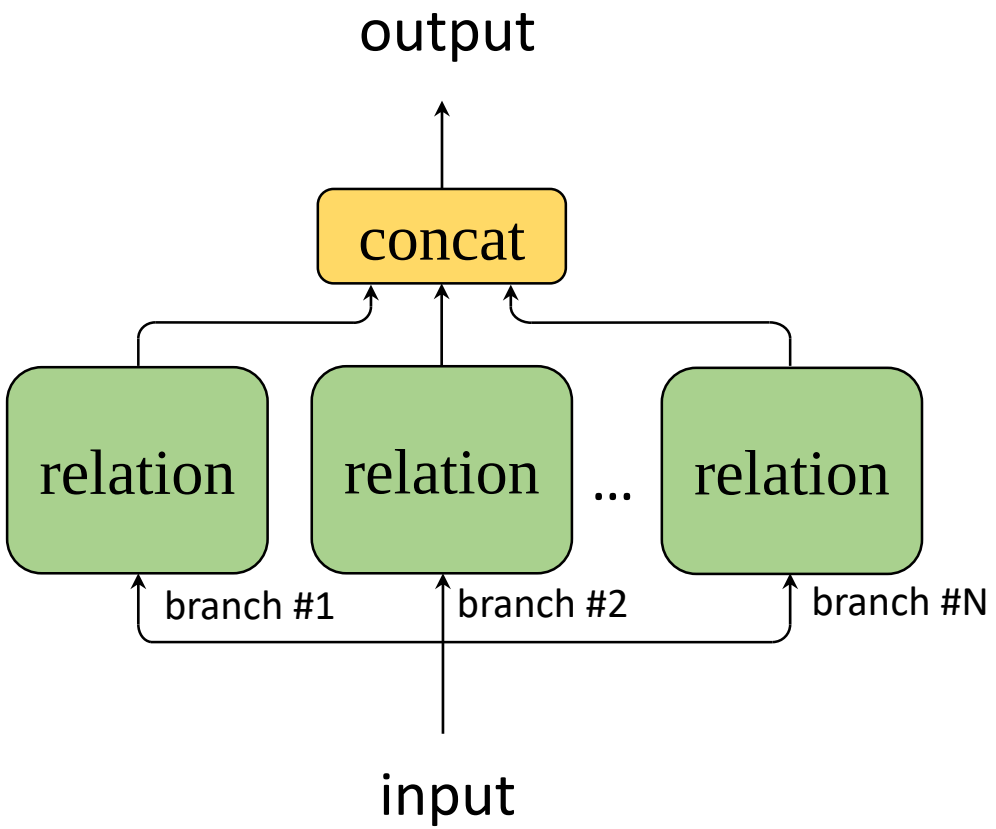
branch #1
(person->glove)



branch #2
(playground->glove)



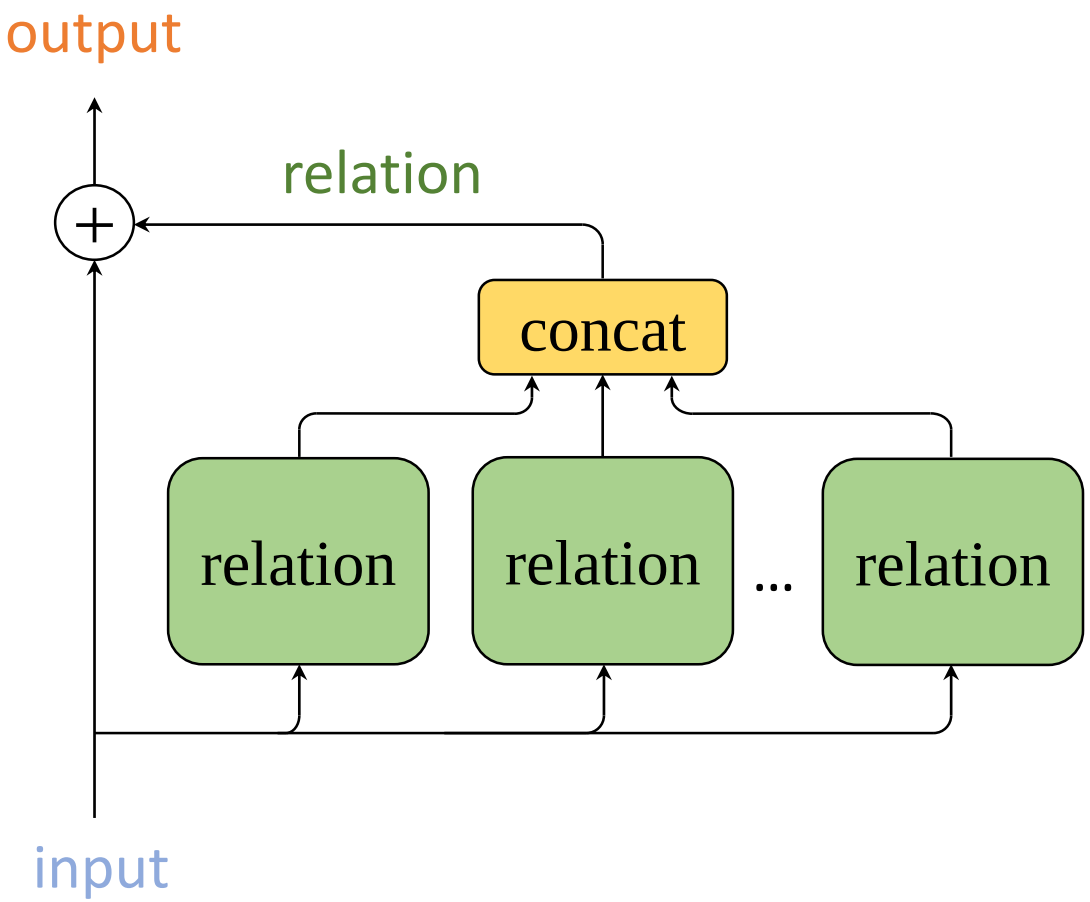
branch #N
(duplicate proposals)



Object Relation Module

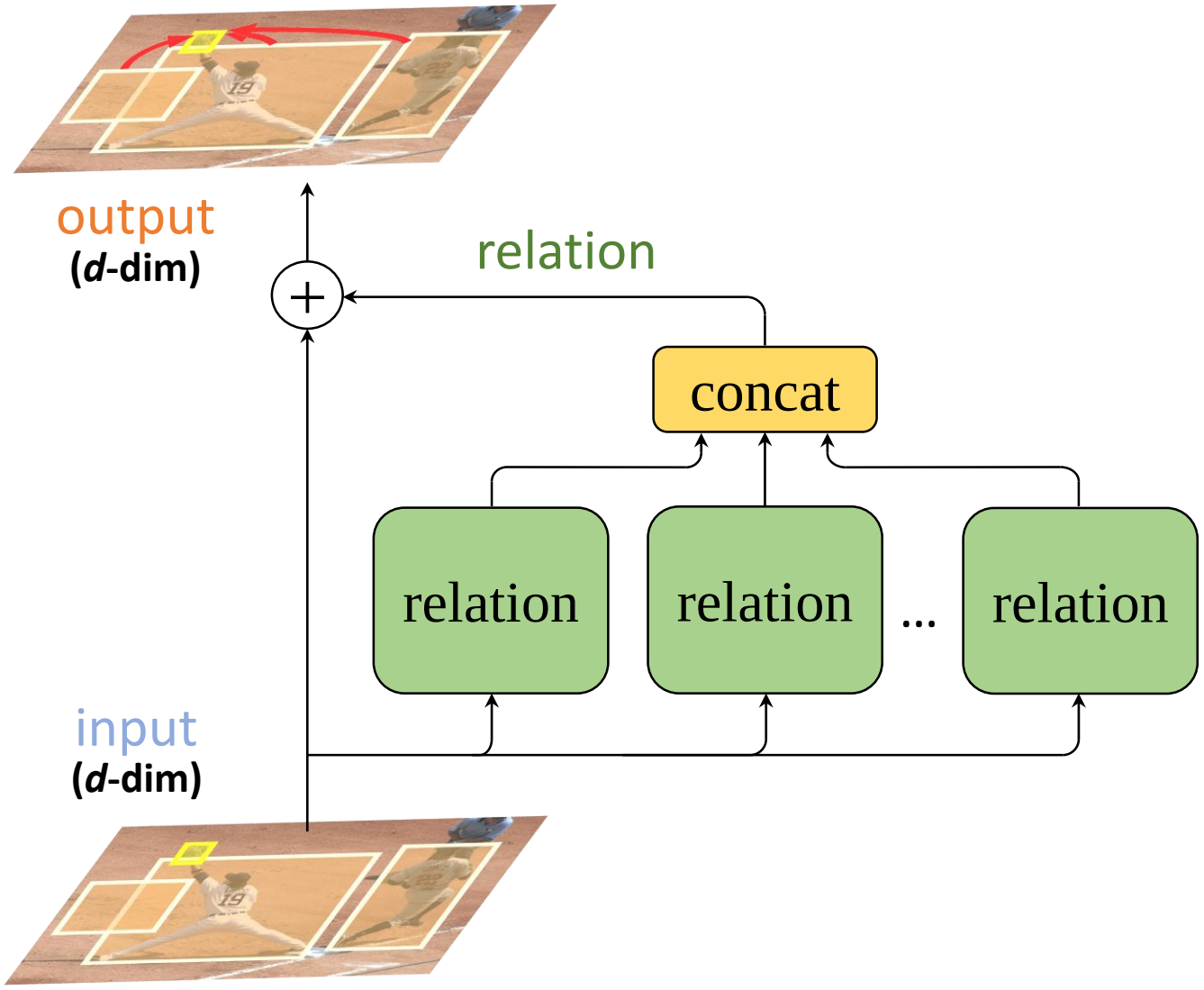
- residual structure

output = input + relation



Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep Residual Learning for Image Recognition. CVPR, 2016

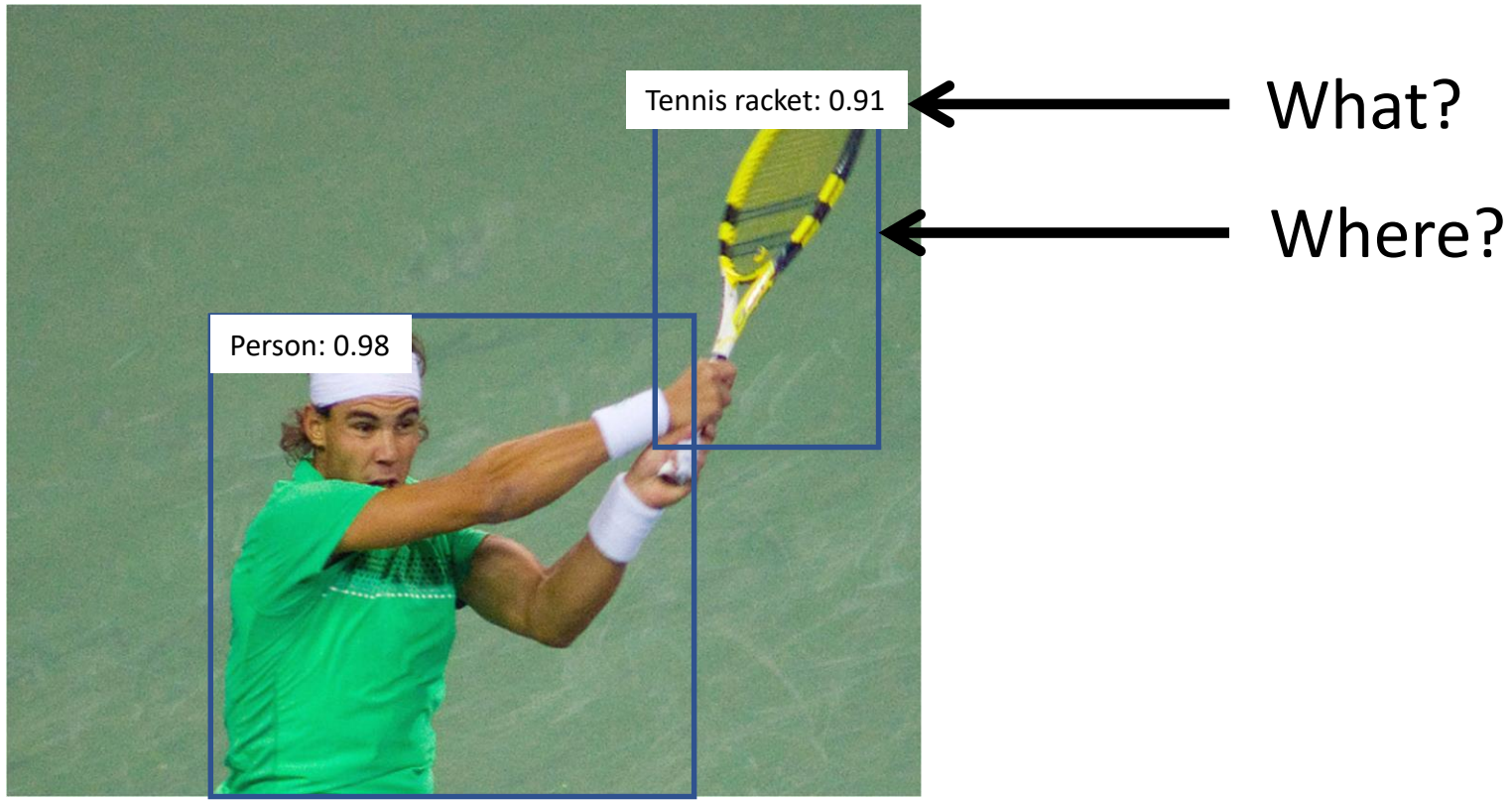
Object Relation Module



Effective and Easy to use

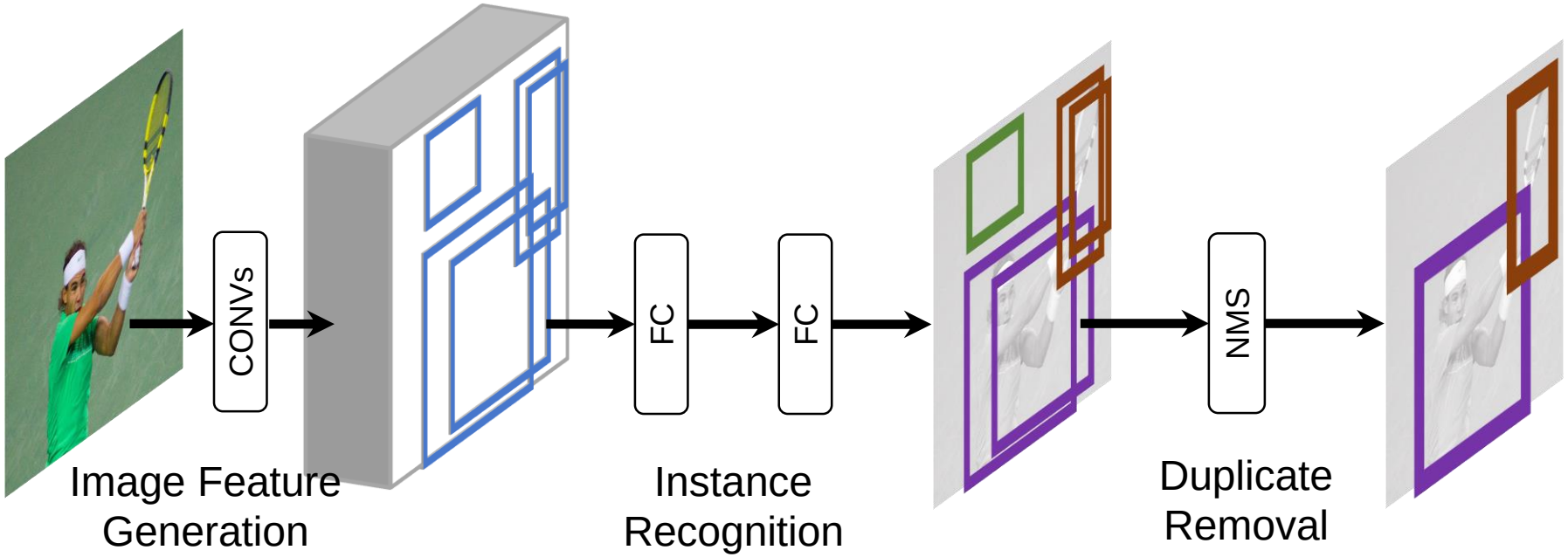
- ✓Parallel
- ✓Learnable
- ✓Require no relation supervision
- ✓Translational invariant
- ✓Stackable

Application: Object Detection



Region-based Object Detection

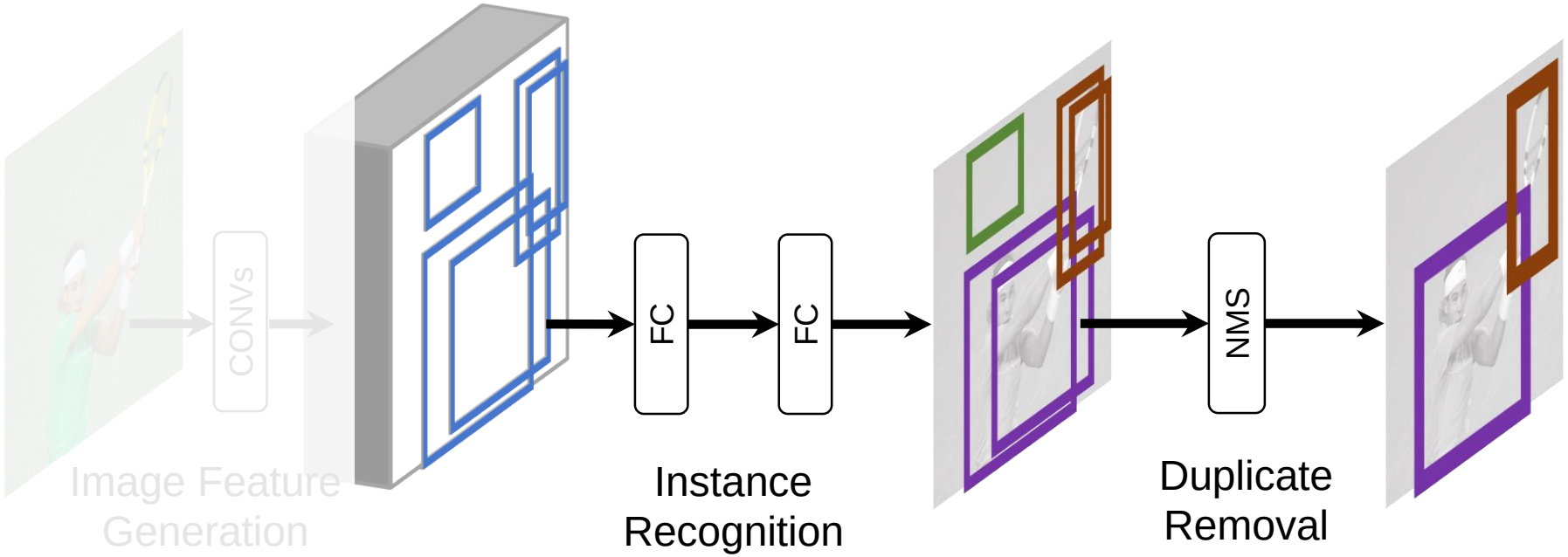
- Fast/Faster R-CNN



R. Girshick. Fast R-CNN. ICCV, 2015
S. Ren et al. Faster R-CNN. NIPS, 2015

Region-based Object Detection

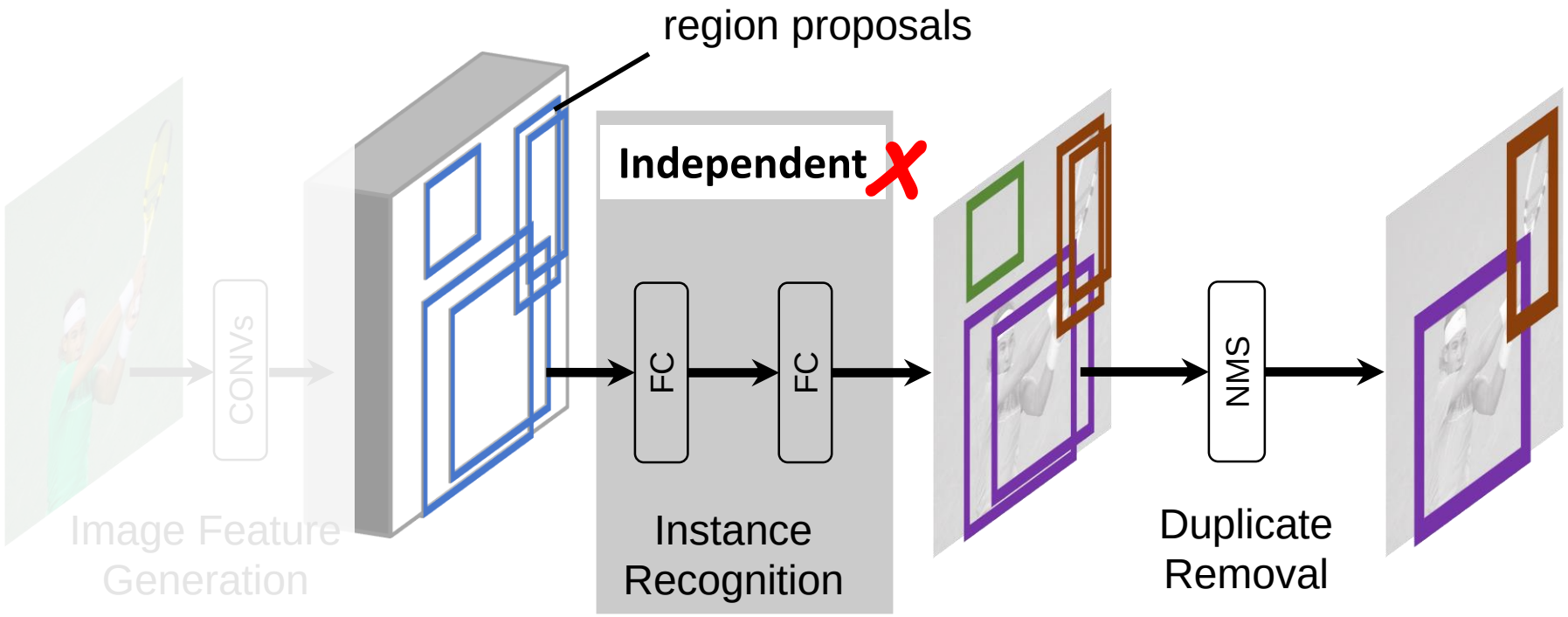
- Fast/Faster R-CNN



R. Girshick. Fast R-CNN. ICCV, 2015
S. Ren et al. Faster R-CNN. NIPS, 2015

Region-based Object Detection

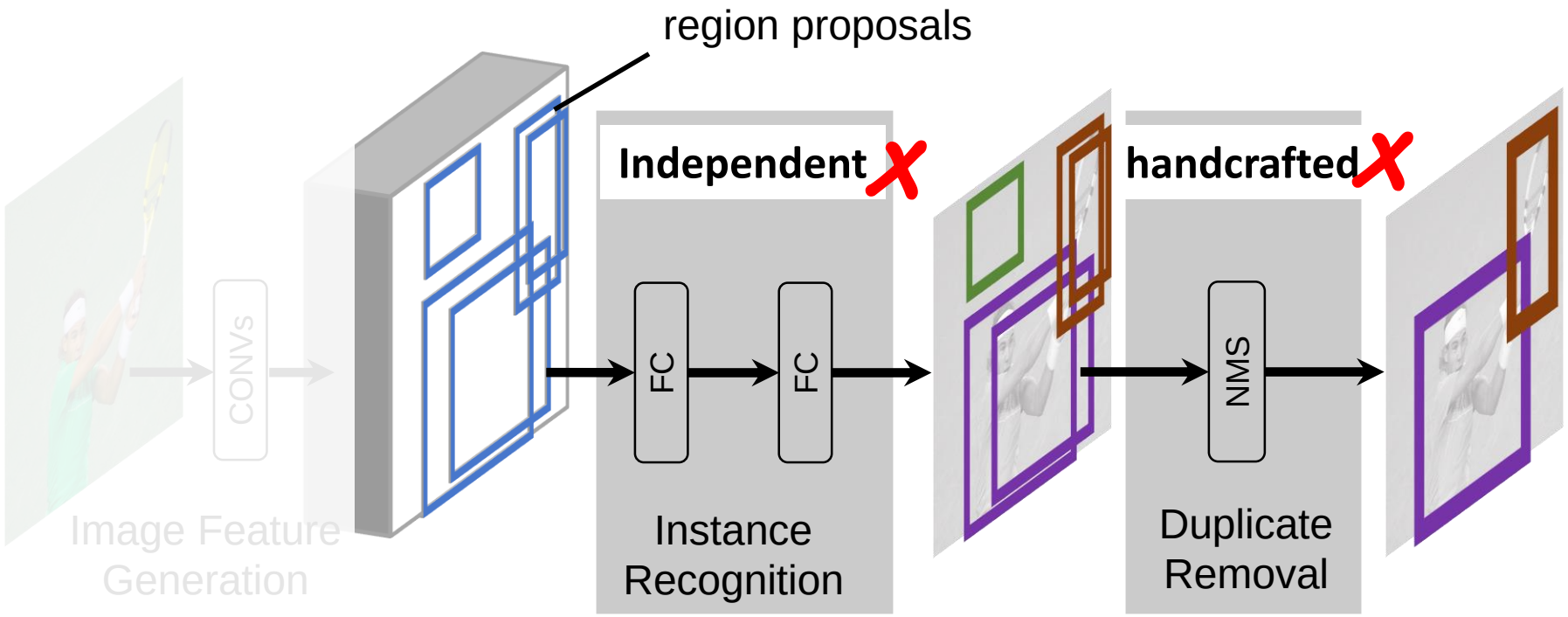
- Fast/Faster R-CNN



R. Girshick. Fast R-CNN. ICCV, 2015
S. Ren et al. Faster R-CNN. NIPS, 2015

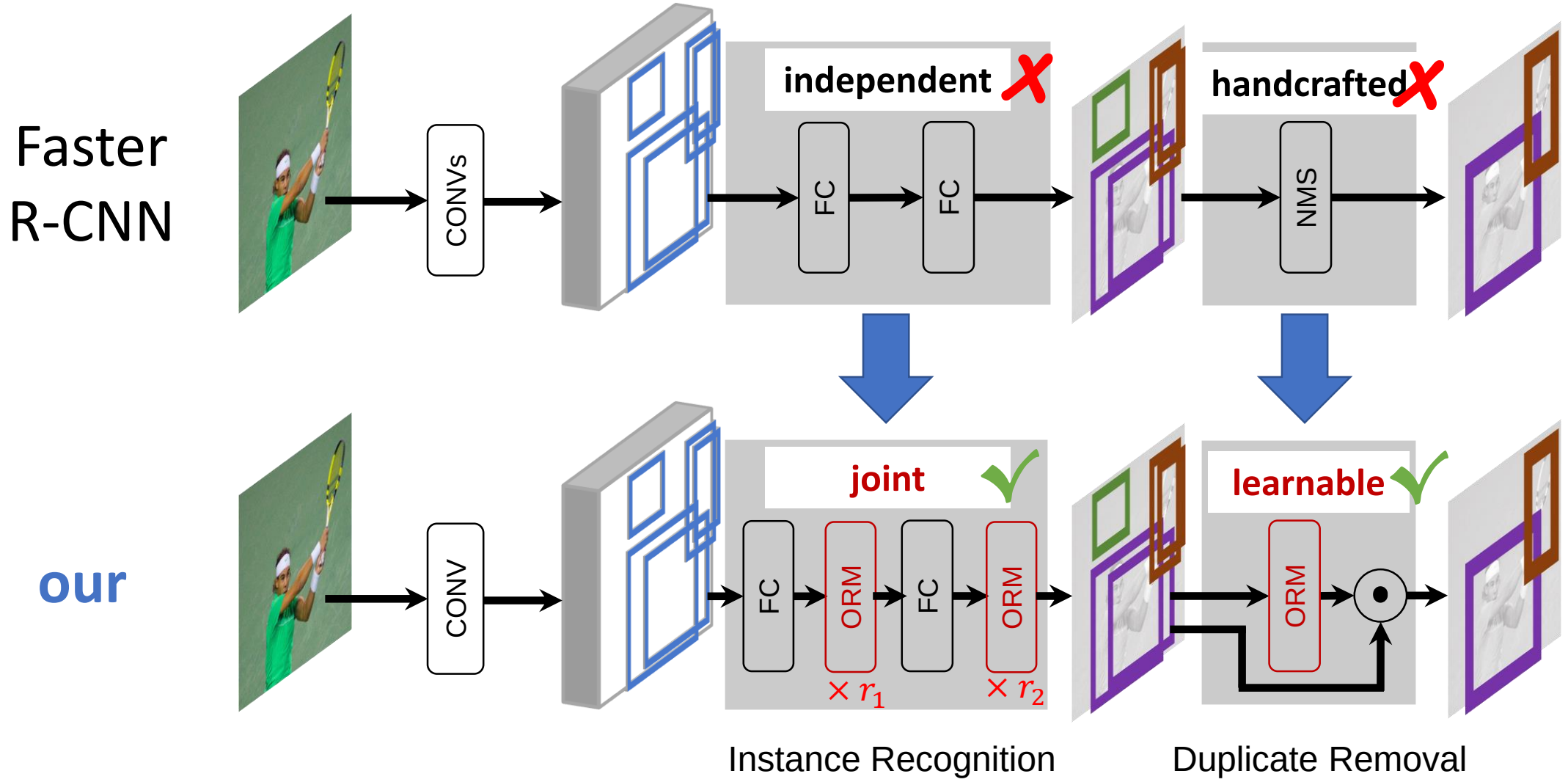
Region-based Object Detection

- Fast/Faster R-CNN

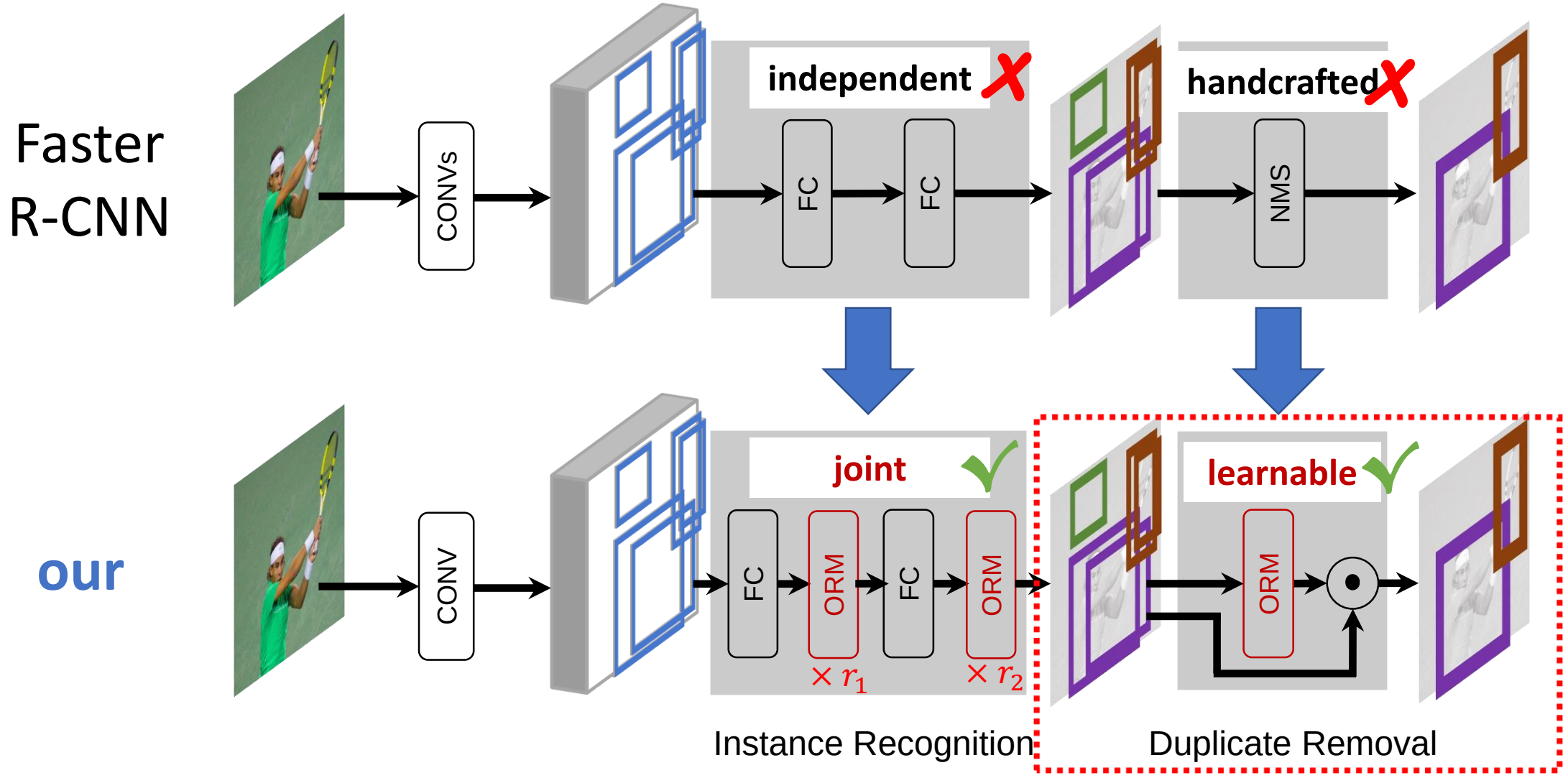


R. Girshick. Fast R-CNN. ICCV, 2015
S. Ren et al. Faster R-CNN. NIPS, 2015

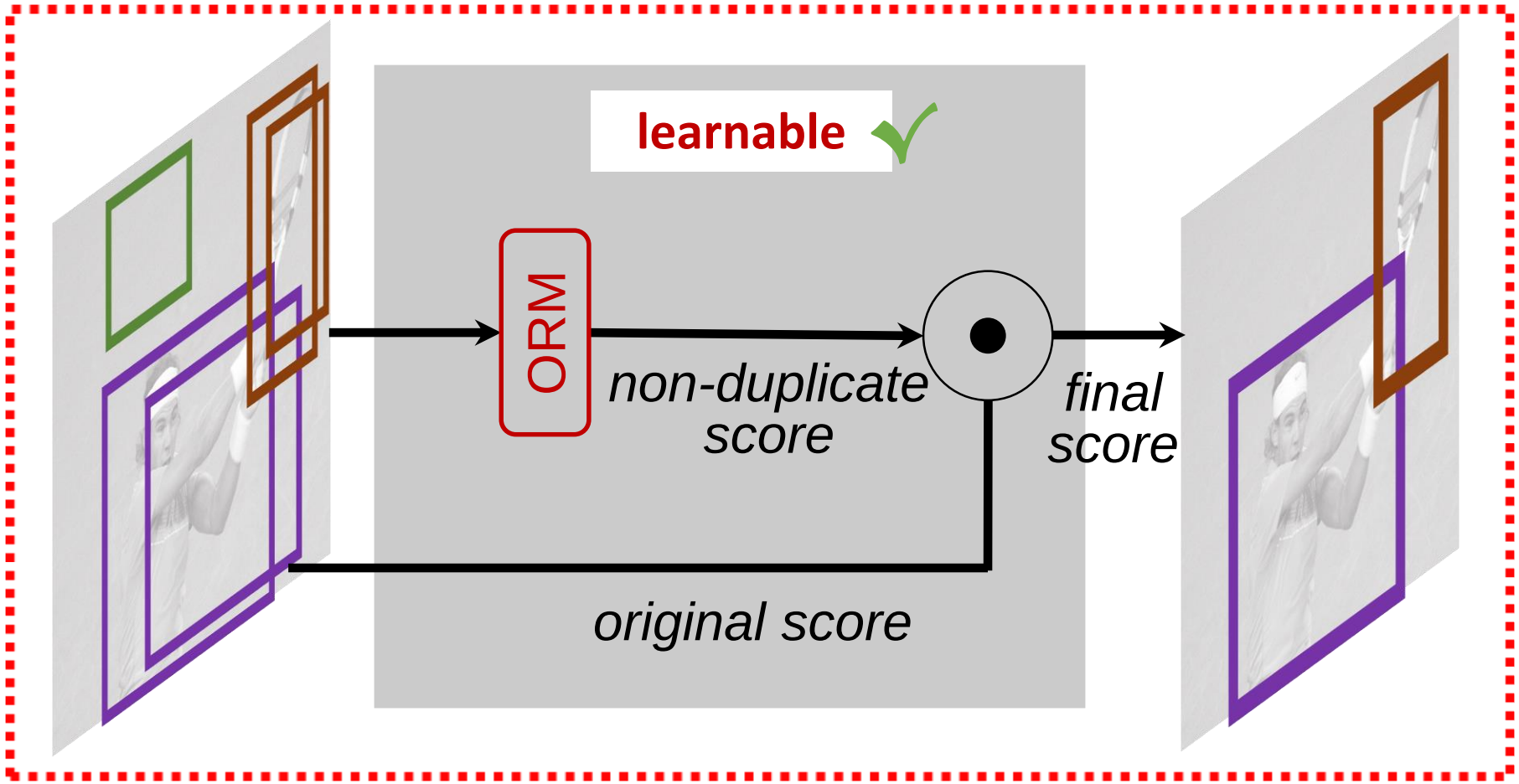
Our method: inserting object relation modules (ORMs)



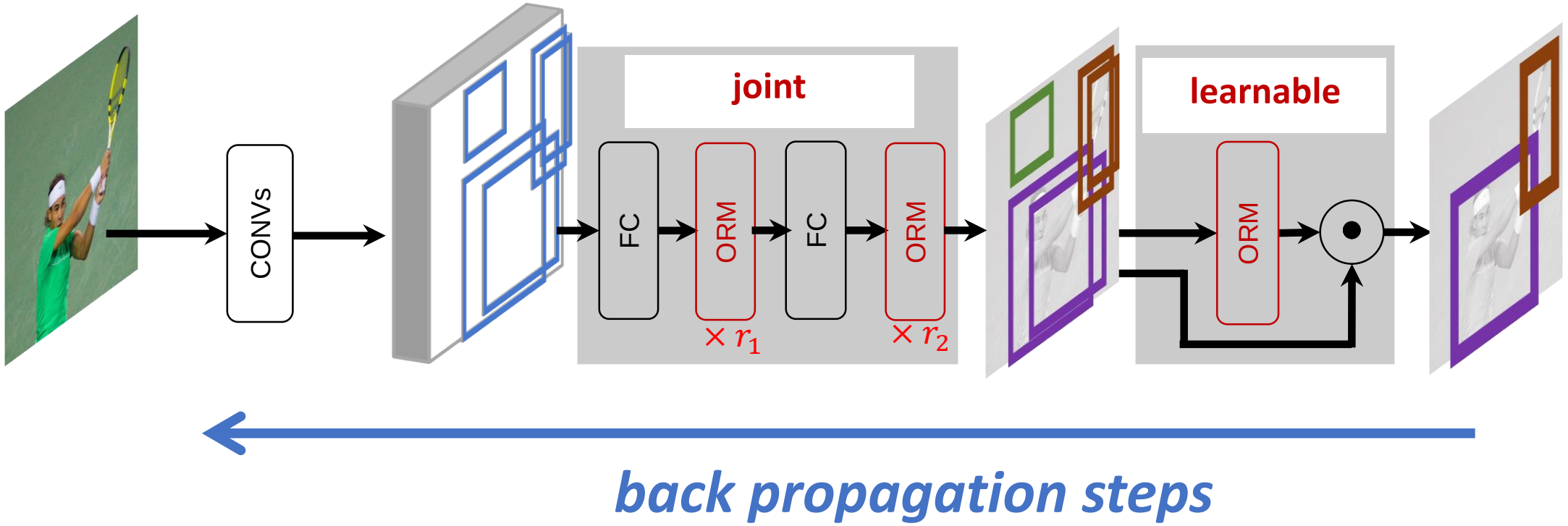
Our method: inserting object relation modules (ORMs)



Learnable Duplicate Removal

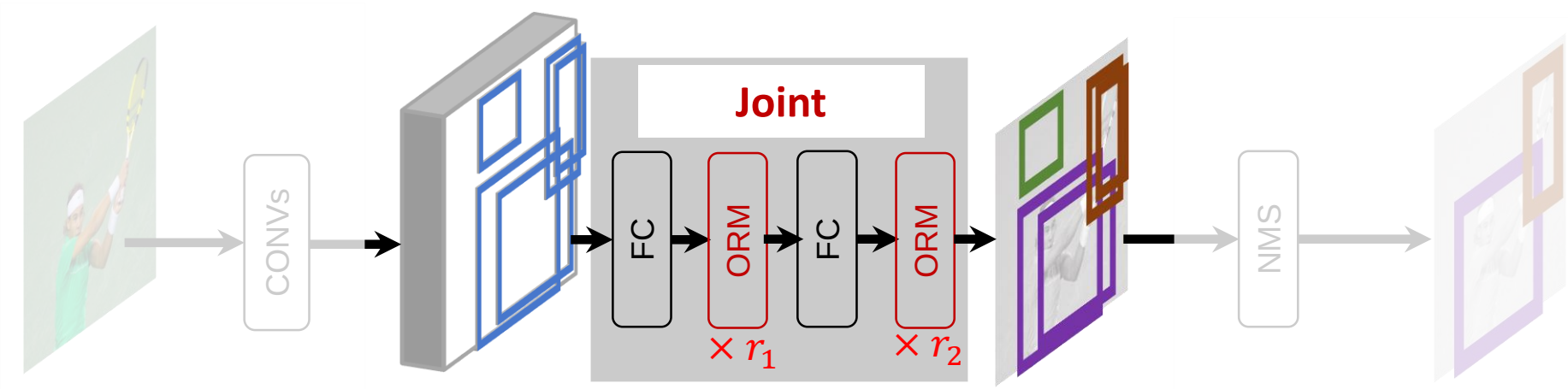


The First Fully End-to-End Object Detector



Results

Instance Recognition Experiments on COCO

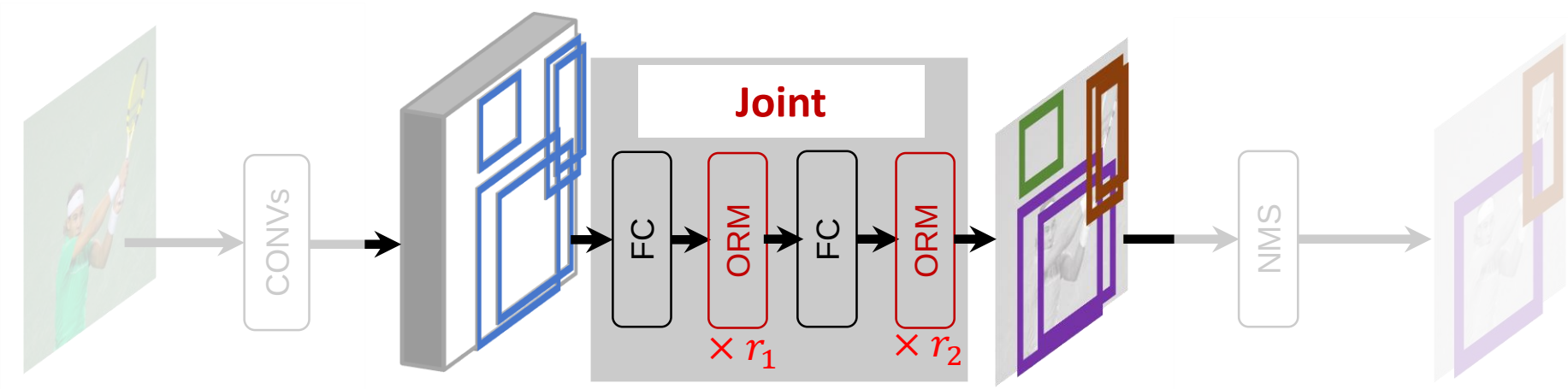


faster R-CNN	+ object relation modules				
	w.o. geometric weight	*w.o.* multi-branch	*w.o.* residual	$\{r_1, r_2\} = \{1, 1\}$	$\{r_1, r_2\} = \{4, 4\}$
29.6	30.3	30.5	30.9	31.9	32.8

*Faster R-CNN with ResNet-50 model are used

- **+2.3** mAP by inserting 2 ORMs
- with +3% FLOPs

Instance Recognition Experiments on COCO

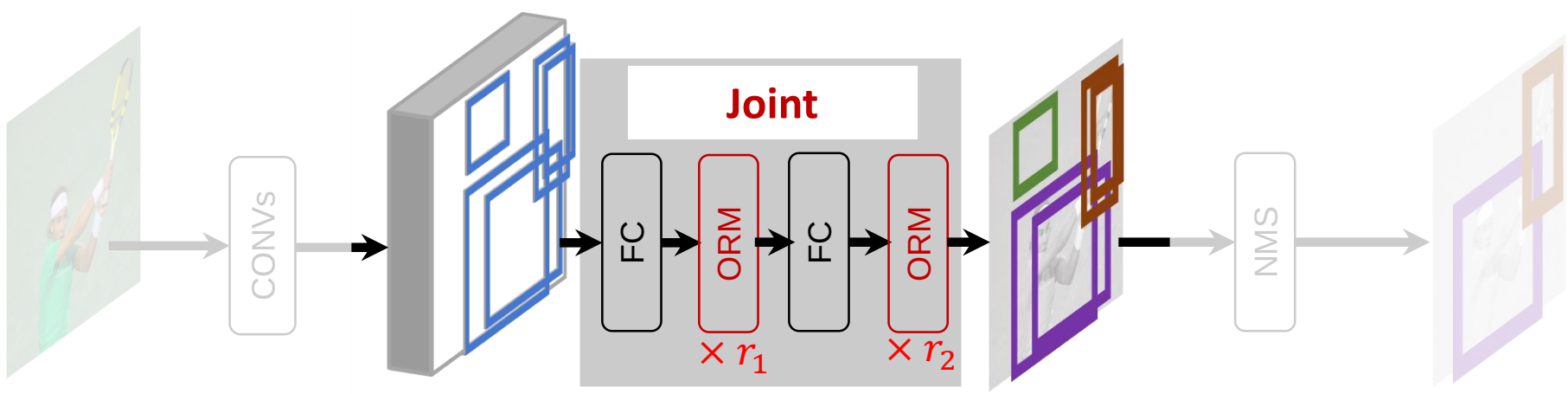


faster R-CNN	+ object relation modules				
	w.o. geometric weight	*w.o.* multi-branch	*w.o.* residual	$\{r_1, r_2\} = \{1, 1\}$	$\{r_1, r_2\} = \{4, 4\}$
29.6	30.3	30.5	30.9	31.9	32.8

*Faster R-CNN with ResNet-50 model are used

- More modules: 8 ORMs

Instance Recognition Experiments on COCO

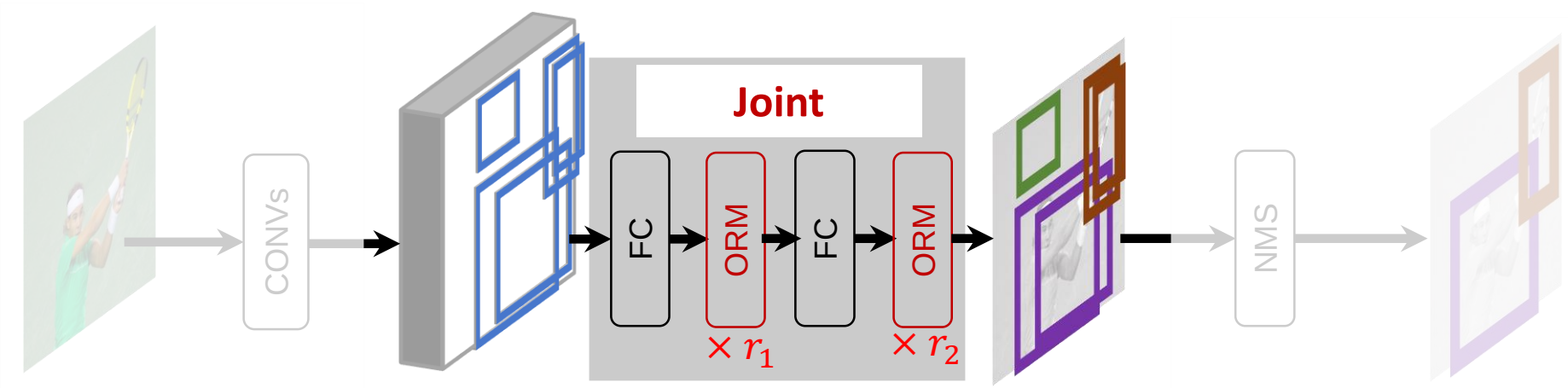


faster R-CNN	+ object relation modules				
	w.o. geometric weight	*w.o.* multi-branch	*w.o.* residual	$\{r_1, r_2\} = \{1, 1\}$	$\{r_1, r_2\} = \{4, 4\}$
29.6	30.3	30.5	30.9	31.9	32.8

*Faster R-CNN with ResNet-50 model are used

- Importance of **relative geometric weight**

Instance Recognition Experiments on COCO

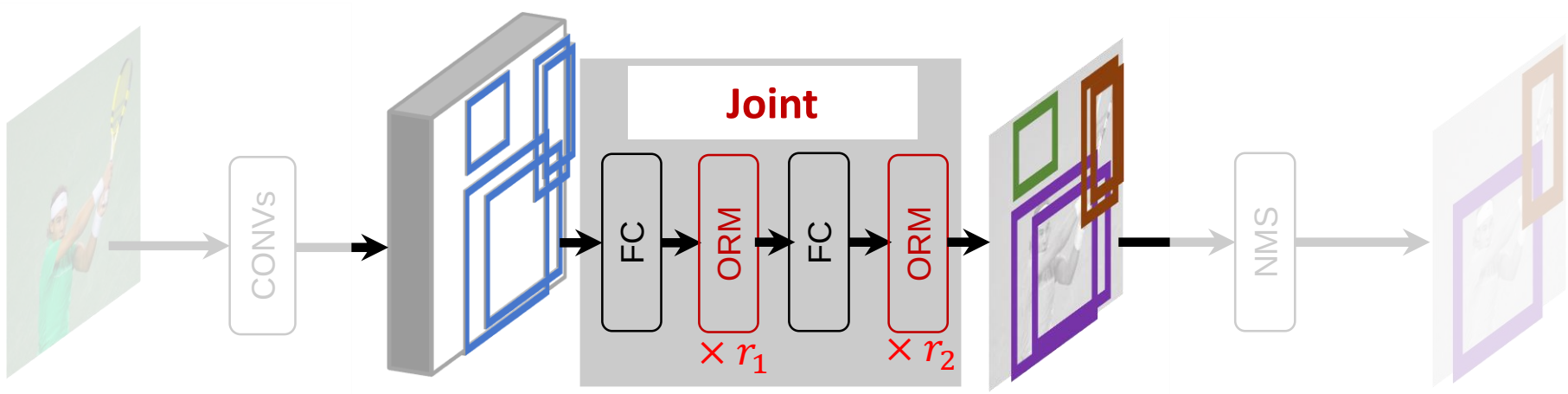


faster R-CNN	+ object relation modules				
	w.o. geometric weight	*w.o.* multi-branch	*w.o.* residual	$\{r_1, r_2\} = \{1, 1\}$	$\{r_1, r_2\} = \{4, 4\}$
29.6	30.3	30.5	30.9	31.9	32.8

*Faster R-CNN with ResNet-50 model are used

- Importance of **multi-branch relation**

Instance Recognition Experiments on COCO

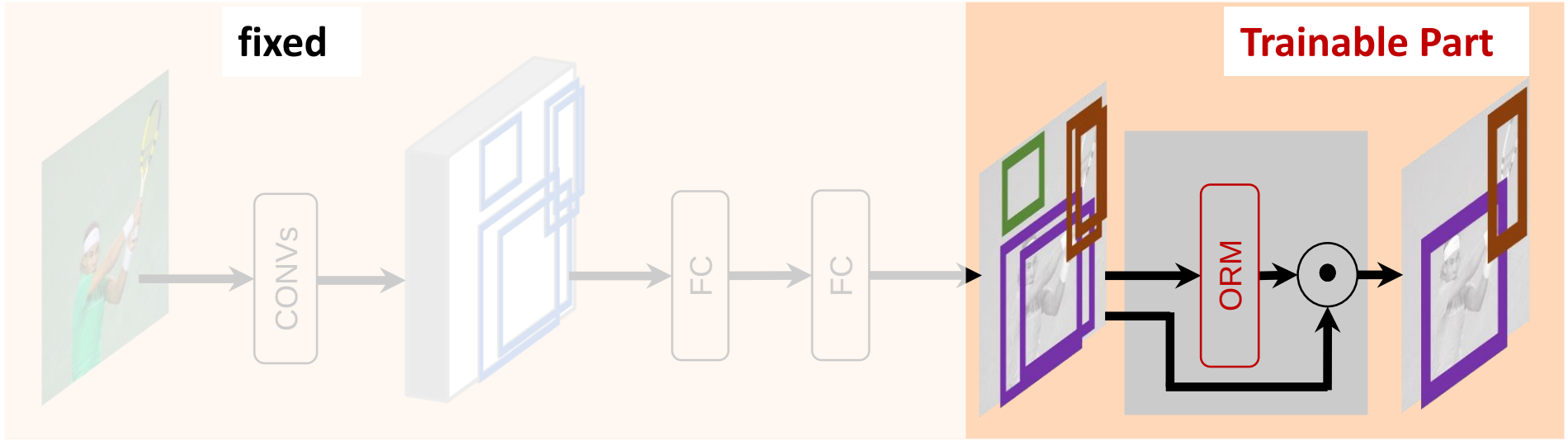


faster R-CNN	+ object relation modules				
	w.o. geometric weight	*w.o.* multi-branch	*w.o.* residual	$\{r_1, r_2\} = \{1, 1\}$	$\{r_1, r_2\} = \{4, 4\}$
29.6	30.3	30.5	30.9	31.9	32.8

*Faster R-CNN with ResNet-50 model are used

- Importance of **residual connection**

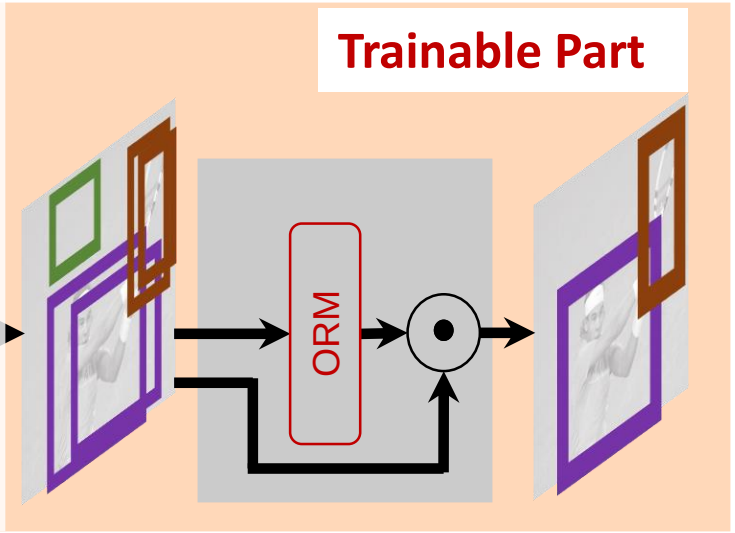
Duplicate Removal Experiments



Duplicate Removal Experiments

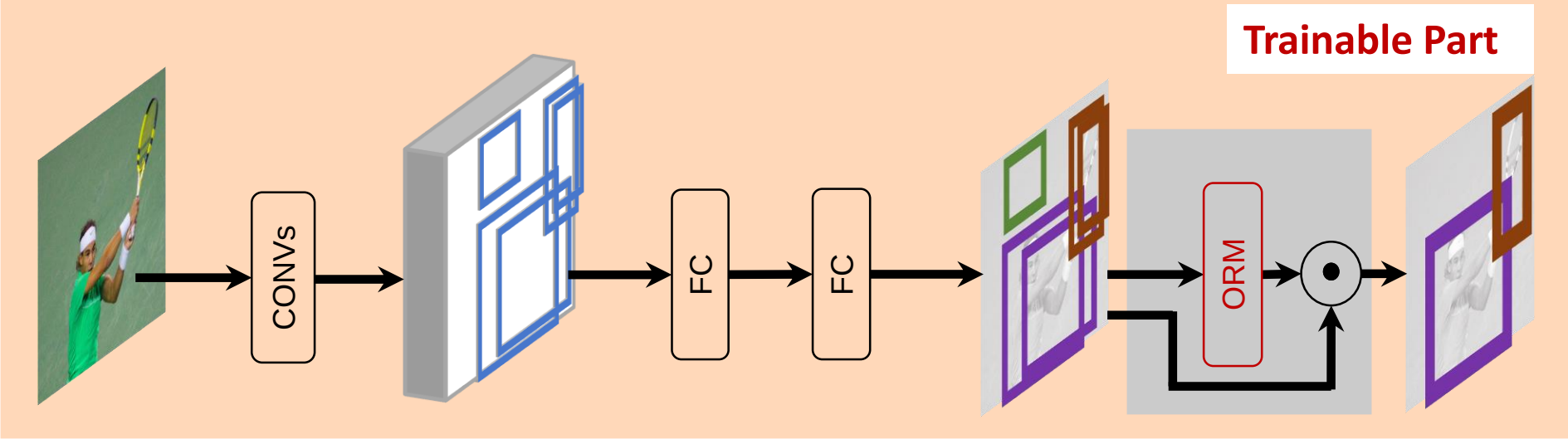
fixed

method	parameters	mAP	mAP_{50}	mAP_{75}
NMS	$N_t = 0.3$	29.0	51.4	29.4
NMS	$N_t = 0.4$	29.4	**52.1**	29.5
NMS	$N_t = 0.5$	29.6	51.9	29.7
NMS	$N_t = 0.6$	**29.6**	50.9	30.1
NMS	$N_t = 0.7$	28.4	46.6	**30.7**
SoftNMS	$\sigma = 0.2$	30.0	**52.3**	30.5
SoftNMS	$\sigma = 0.4$	30.2	51.7	31.3
SoftNMS	$\sigma = 0.6$	**30.2**	50.9	31.6
SoftNMS	$\sigma = 0.8$	29.9	49.9	**31.6**
SoftNMS	$\sigma = 1.0$	29.7	49.7	31.6
ours	$\eta = 0.5$	30.3	**51.9**	31.5
ours	$\eta = 0.75$	30.1	49.0	**32.7**
ours	$\eta \in [0.5, 0.9]$	**30.5**	50.2	32.4



- **Noticeably** better than NMS
- **Slightly** better than SoftNMS [N. Bodla et al, 2017]

Fully End-to-End Object Detection



method	parameters	mAP	mAP_{50}	mAP_{75}
NMS	$N_t = 0.6$	**29.6**	50.9	30.1
SoftNMS	$\sigma = 0.6$	**30.2**	50.9	31.6
ours	$\eta = 0.5$	30.3	**51.9**	31.5
ours	$\eta = 0.75$	30.1	49.0	**32.7**
ours	$\eta \in [0.5, 0.9]$	**30.5**	50.2	32.4
ours (e2e)	$\eta \in [0.5, 0.9]$	**31.0**	51.4	32.8

- Benefit from fully end-to-end training

Using Stronger Backbones

backbone	setting	mAP	mAP_{50}	mAP_{75}	#. params	FLOPS	
faster RCNN	2fc+SoftNMS	32.2/32.7	52.9/53.6	34.2/34.7	58.3M	122.2B	
	2fc+RM+SoftNMS	34.7/35.2	55.3/**56.2**	37.2/37.8	64.3M	124.6B	**+3.0 mAP**
	2fc+RM+e2e	**35.2/35.4**	**55.8**/56.1	**38.2/38.5**	64.6M	124.9B	
FPN	2fc+SoftNMS	36.8/37.2	57.8/58.2	40.7/41.4	56.4M	145.8B	
	2fc+RM+SoftNMS	38.1/38.3	59.5/59.9	41.8/42.3	62.4M	157.8B	**+2.0 mAP**
	2fc+RM+e2e	**38.8/38.9**	**60.3/60.5**	**42.9/43.3**	62.8M	158.2B	
DCN	2fc+SoftNMS	37.5/38.1	57.3/58.1	41.0/41.6	60.5M	125.0B	
	2fc+RM+SoftNMS	38.1/38.8	57.8/**58.7**	41.3/42.4	66.5M	127.4B	**+1.0 mAP**
	2fc+RM+e2e	**38.5/39.0**	**57.8**/58.6	**42.0/42.9**	66.8M	127.7B	

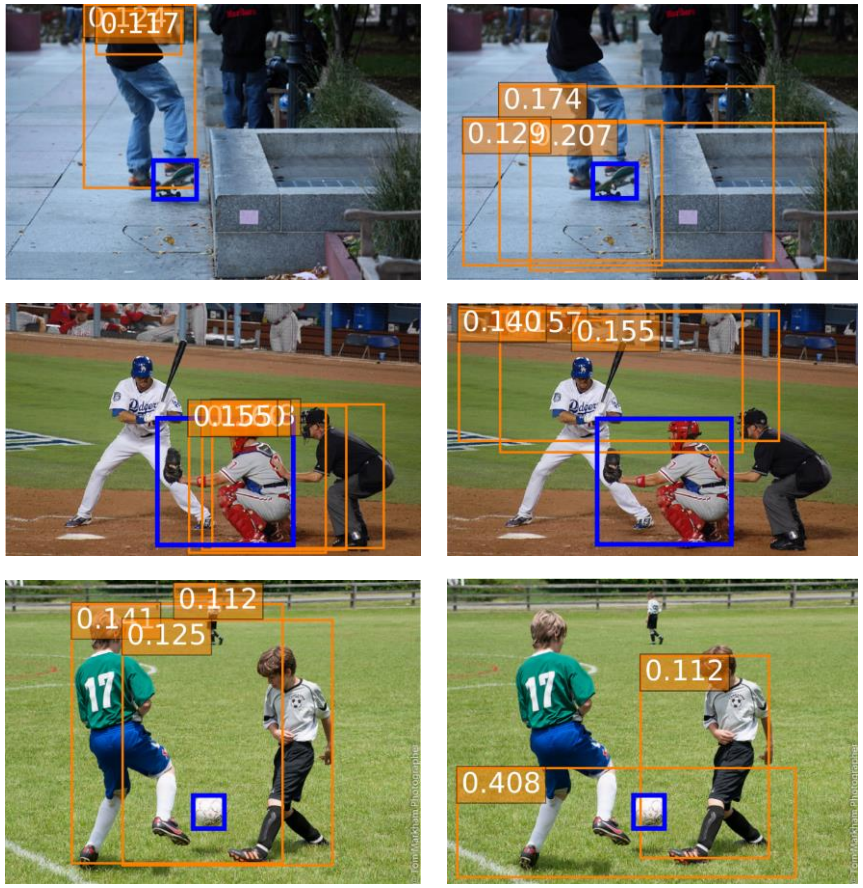
*Faster R-CNN with ResNet-101 model are used (evaluation on *minival*/*test-dev* are reported)

- less than 10% computation overhead on all backbones

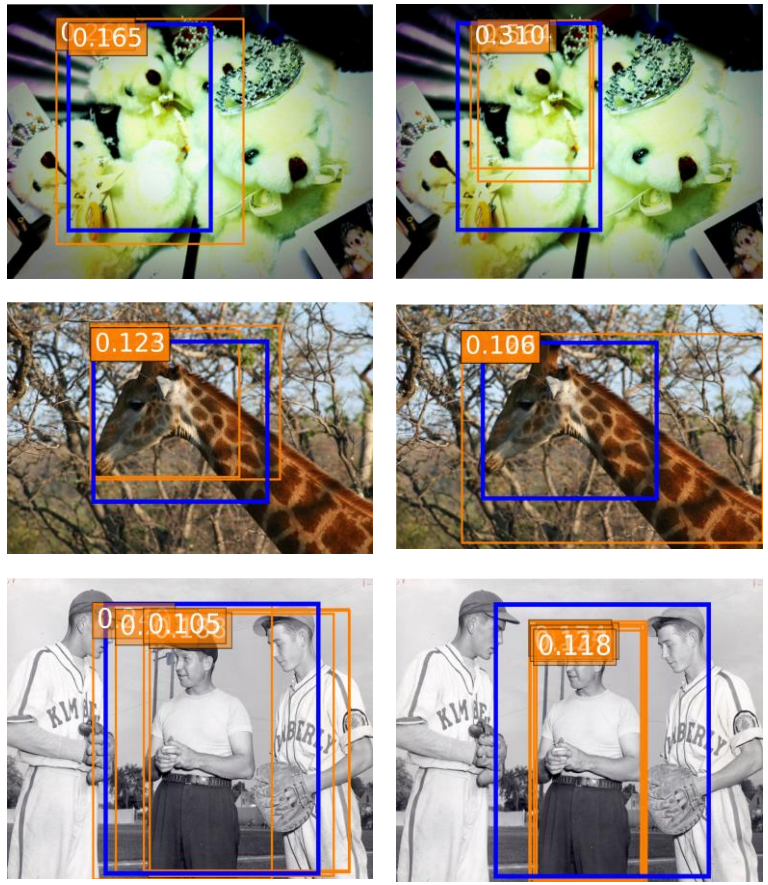
What is Learnt?

Object Pairs with High Relation Weights

instance recognition



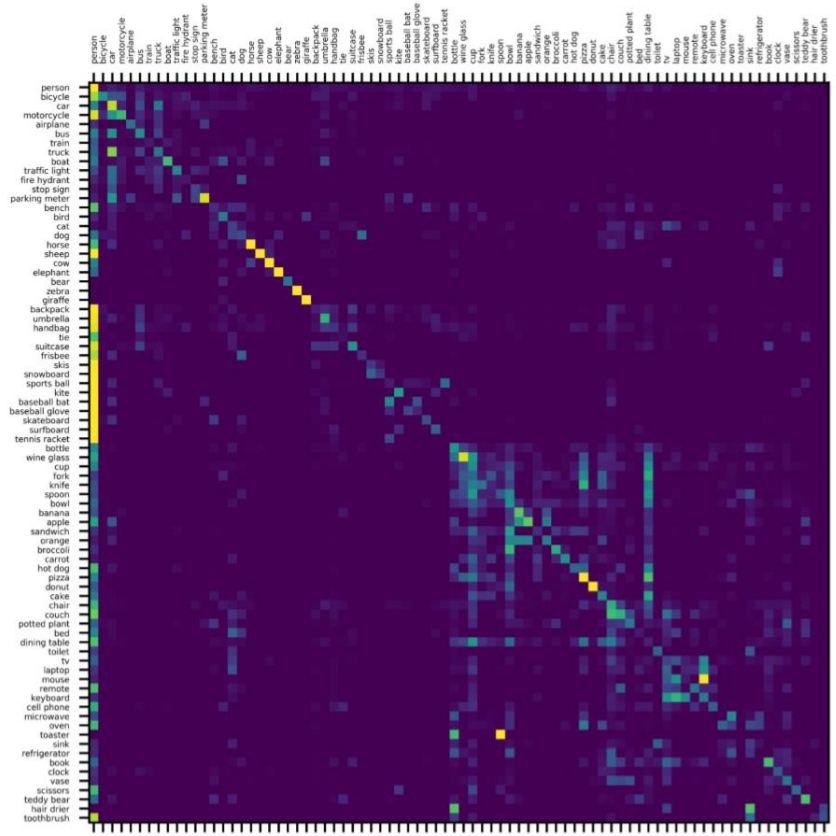
duplicate removal



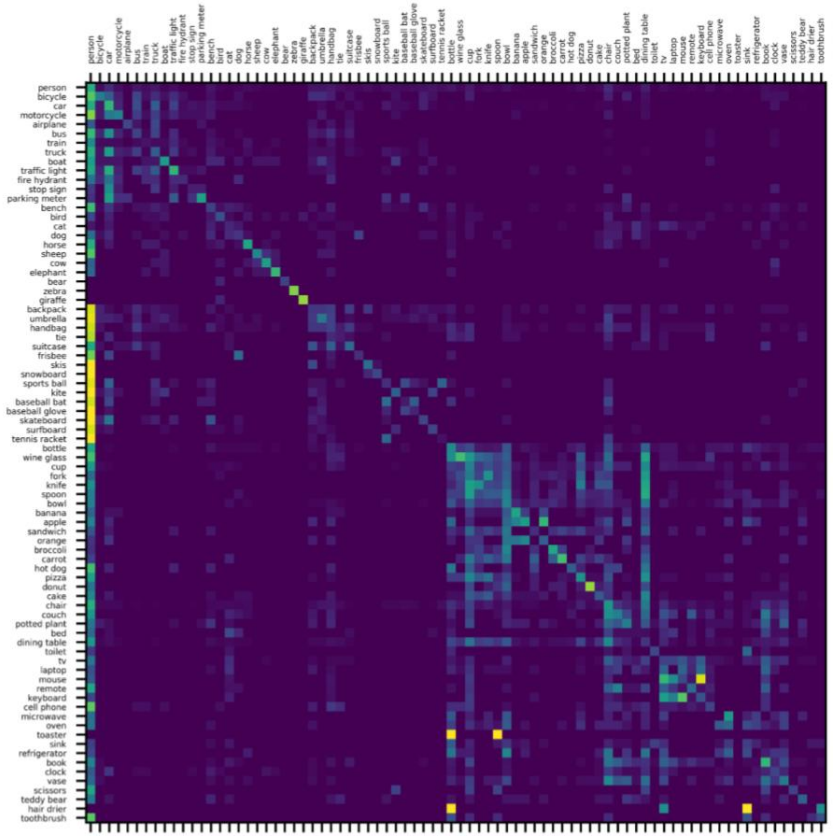
reference object

other objects contributing high weights

Class Co-Occurrence Information is Learnt



Class Co-occurrence Probability

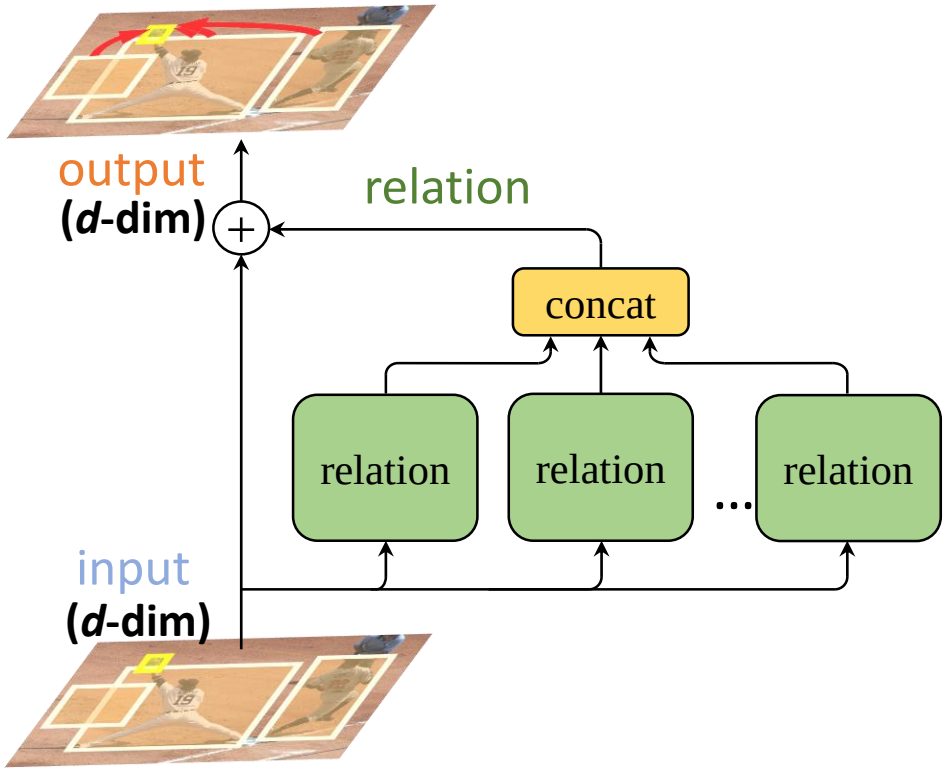


Learnt Attentional Weights

$r = 0.90$

Conclusion

- A novel object relation module to model object-object relation
  - ✓Parallel
  - ✓Learnable
  - ✓Require no relation supervision
  - ✓Translational invariant
  - ✓Stackable
- Application: Object Detection
  - ✓Improves object detection accuracy
  - ✓The first fully end-to-end object detector



code:

https://github.com/msracver/Relation-Networks-for-Object-Detection