Relation Networks for Visual Modeling

Han Hu
Visual Computing Group
Microsoft Research Asia (MSRA)

https://ancientmooner.github.io/

Collaborators: Zheng Zhang, Yue Cao, Jiayuan Gu, Jiarui Xu, Jifeng Dai, Yichen Wei, Stephen Lin, Liwei Wang, Zhenda Xie, Fangyun Wei

Human Brain

- Human cortex can universally perceive different senses

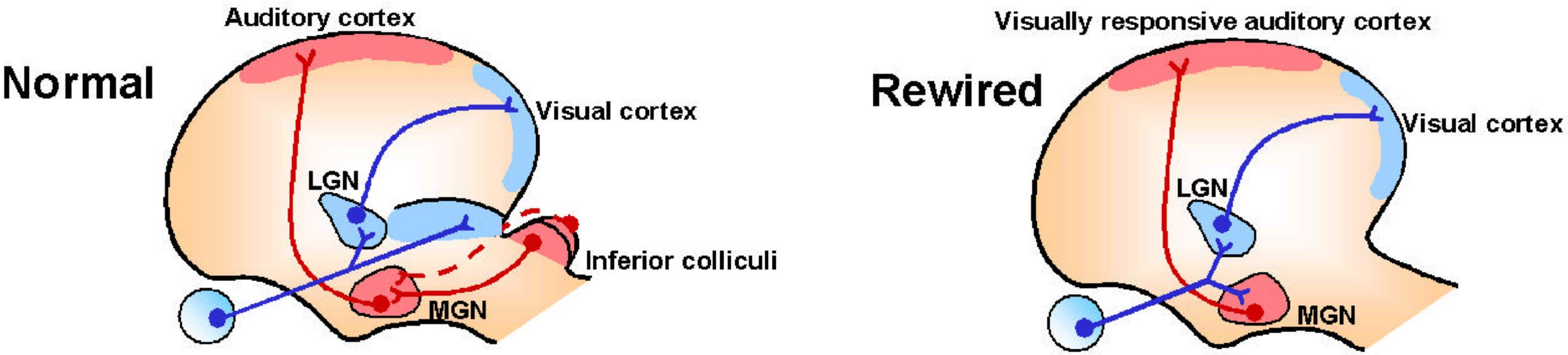
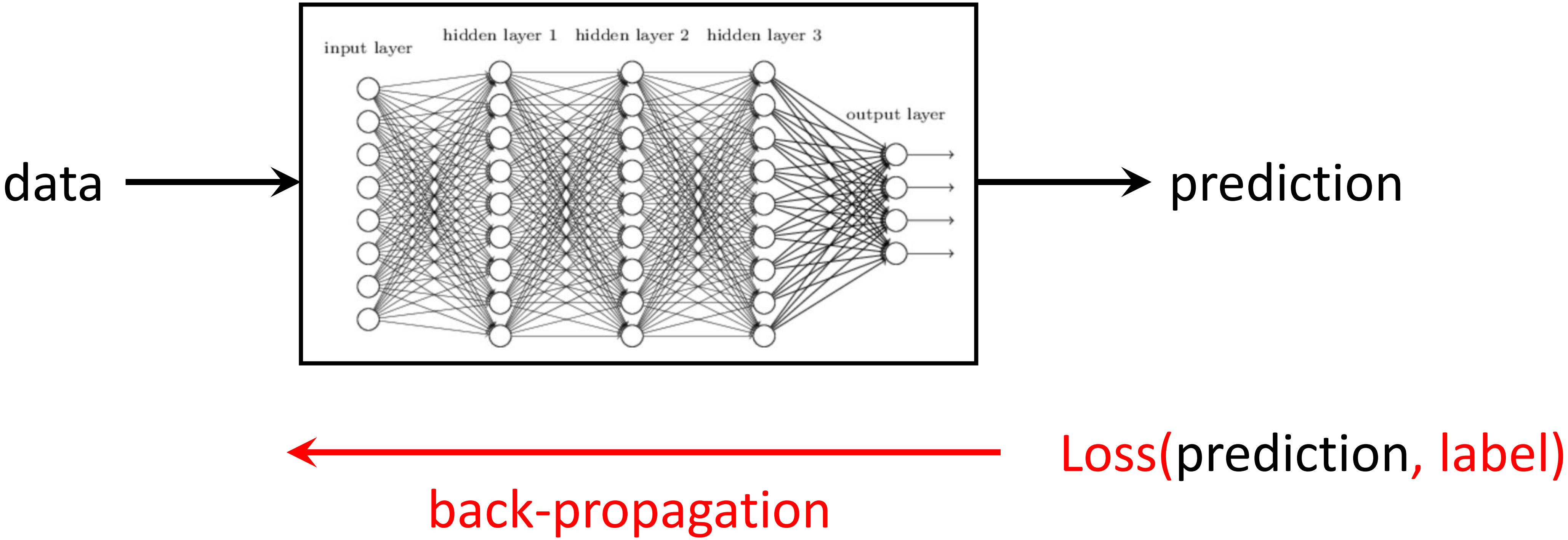


figure credit to J. Sharma et al.

Intelligent Machines

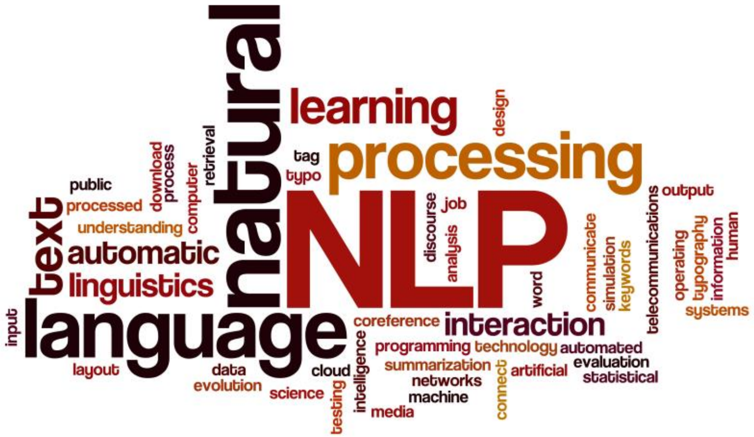
- A **universal** learning pipeline



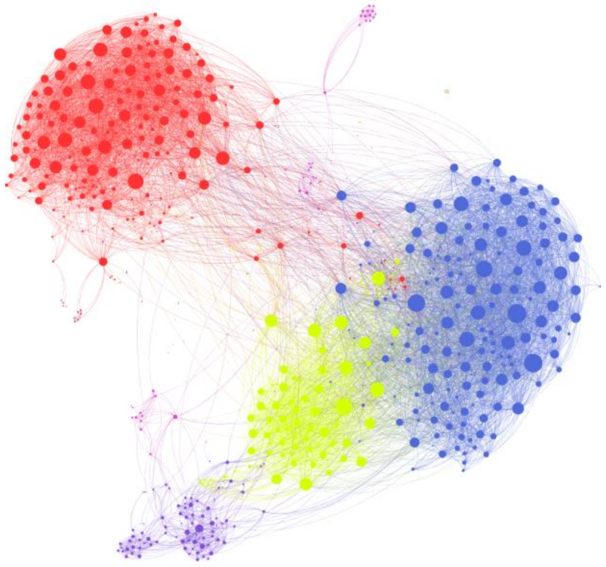
Intelligent Machines

- **Particular** basic model for different task/data

convolution



LSTM, GRU, convolution, self-attention, …

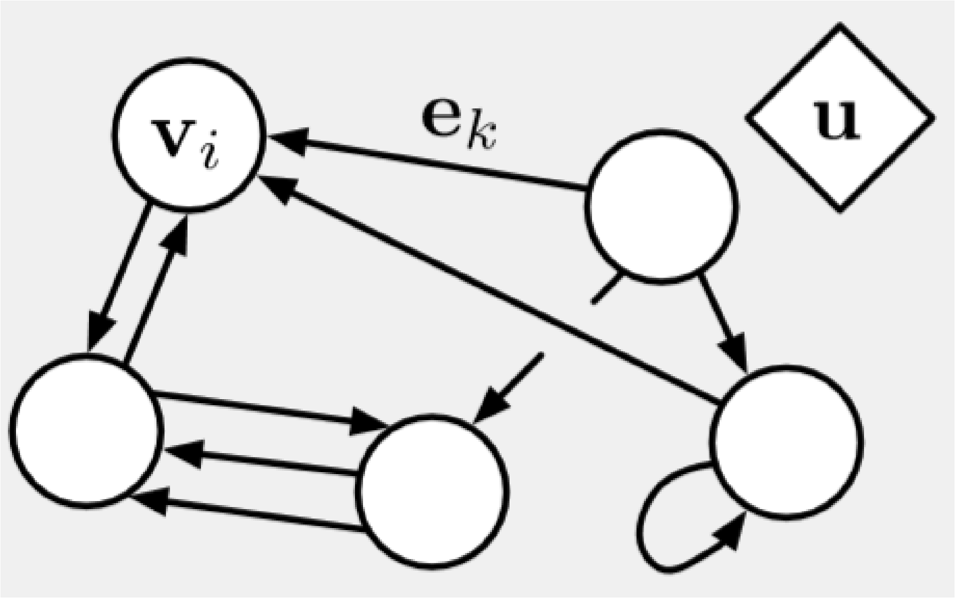


graph networks

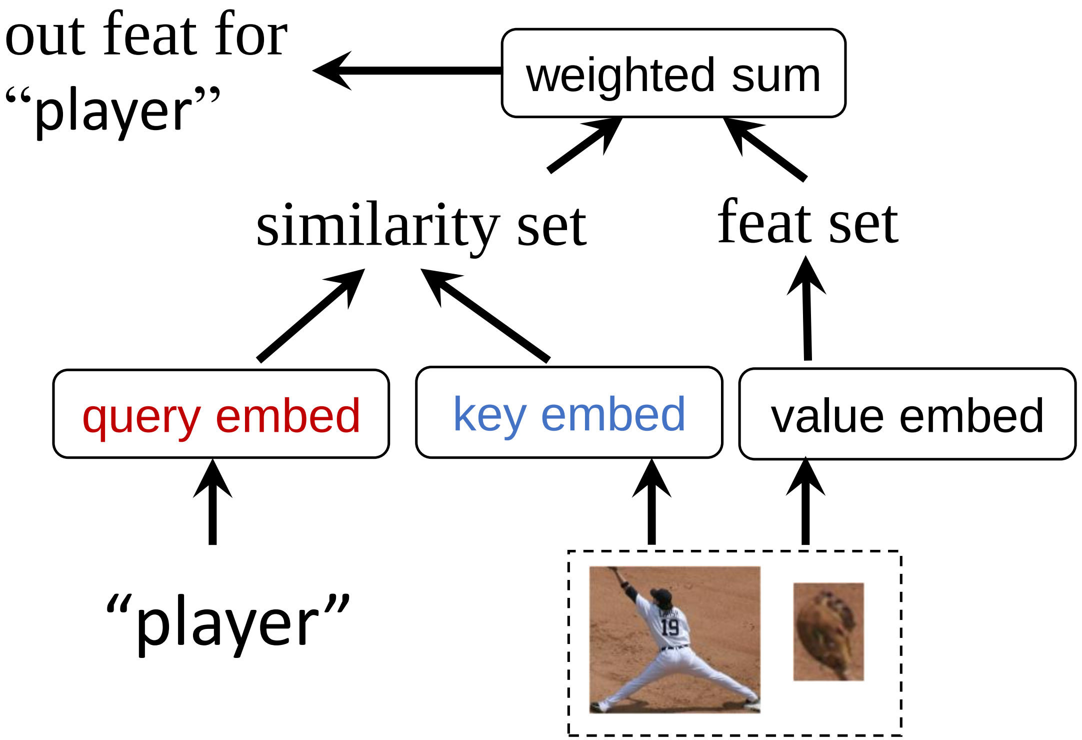
Universal Basic Models for Intelligent Machines?

Relation Networks: Towards Universal Basic Models

similar things: ***graph neural networks***, *self-attention*, …



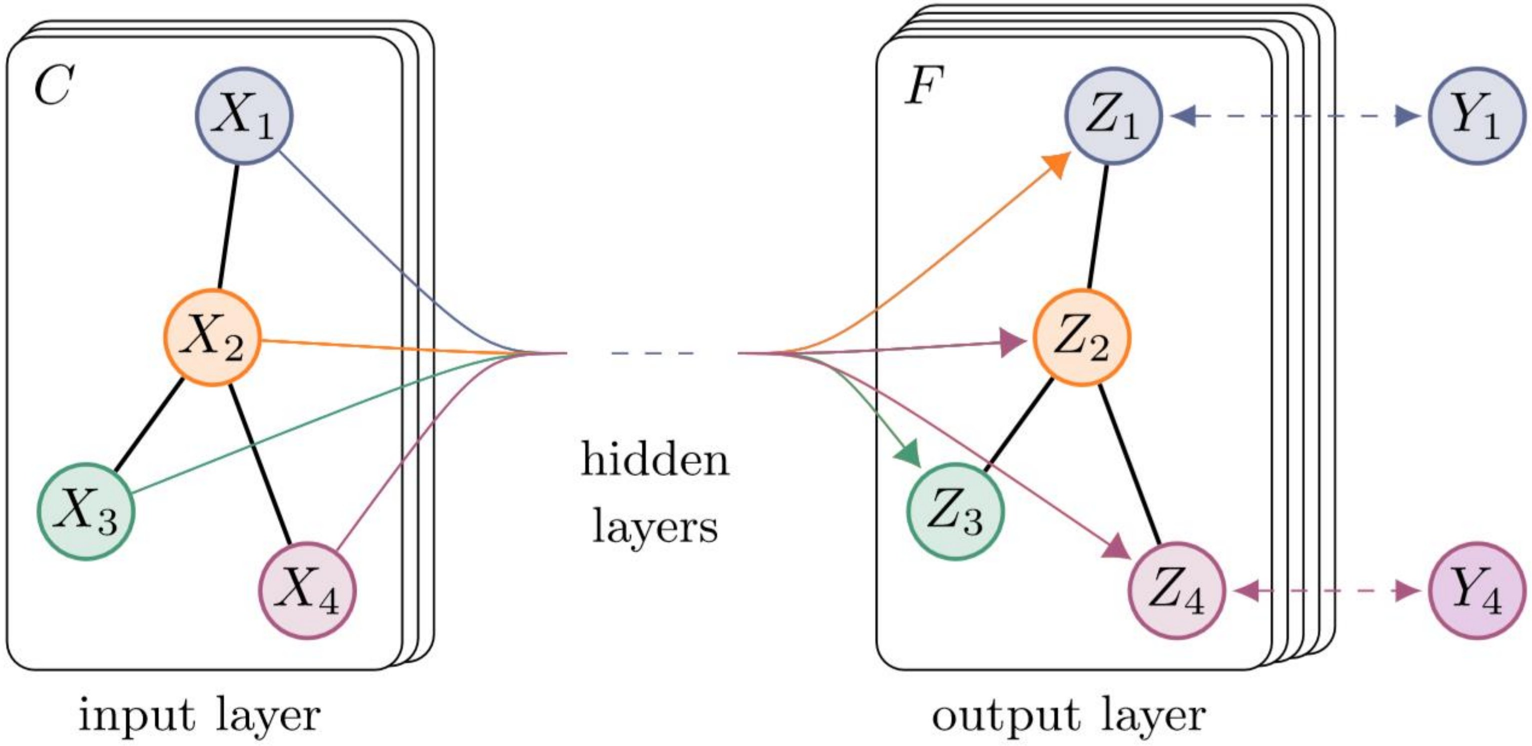
graph neural networks



(self)-attention

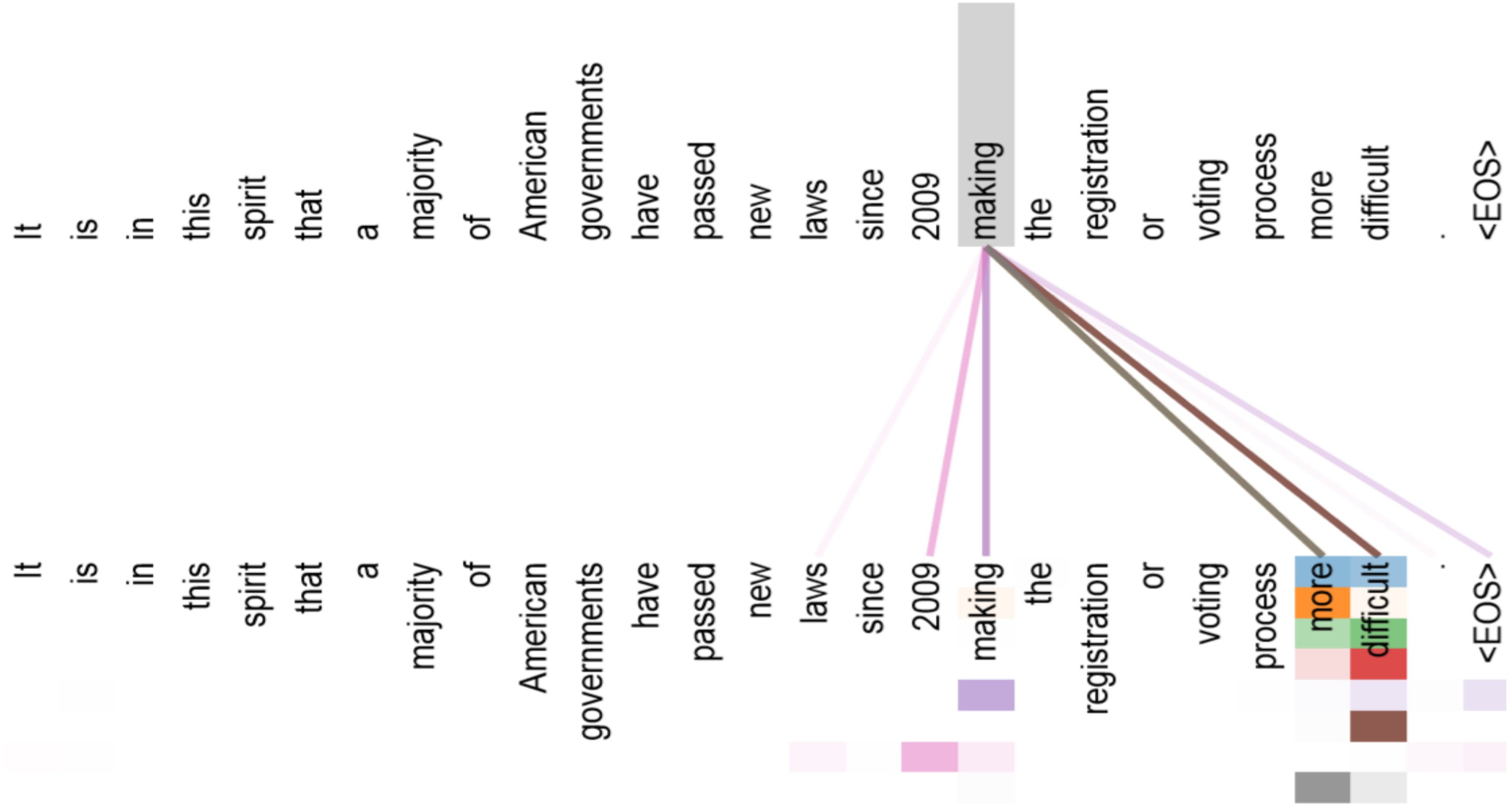
left figure credit to P. Battaglia et al.

Relation Networks for Graph Data



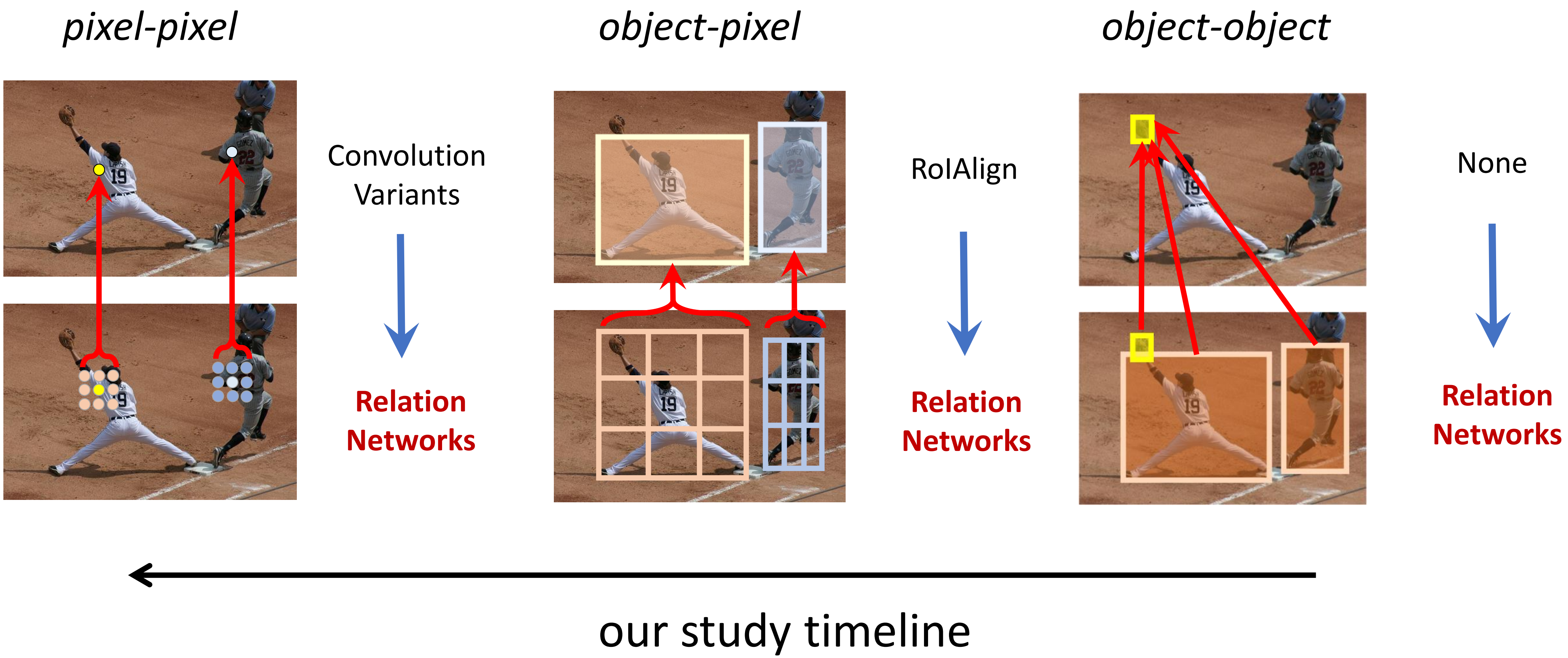
T. Kipf and M. Welling. *Semi-supervised classification with graph convolutional networks.* ICLR 2018

Relation Networks for NLP



A. Vaswani and et al. *Attention Is All You Need.* NIPS 2017

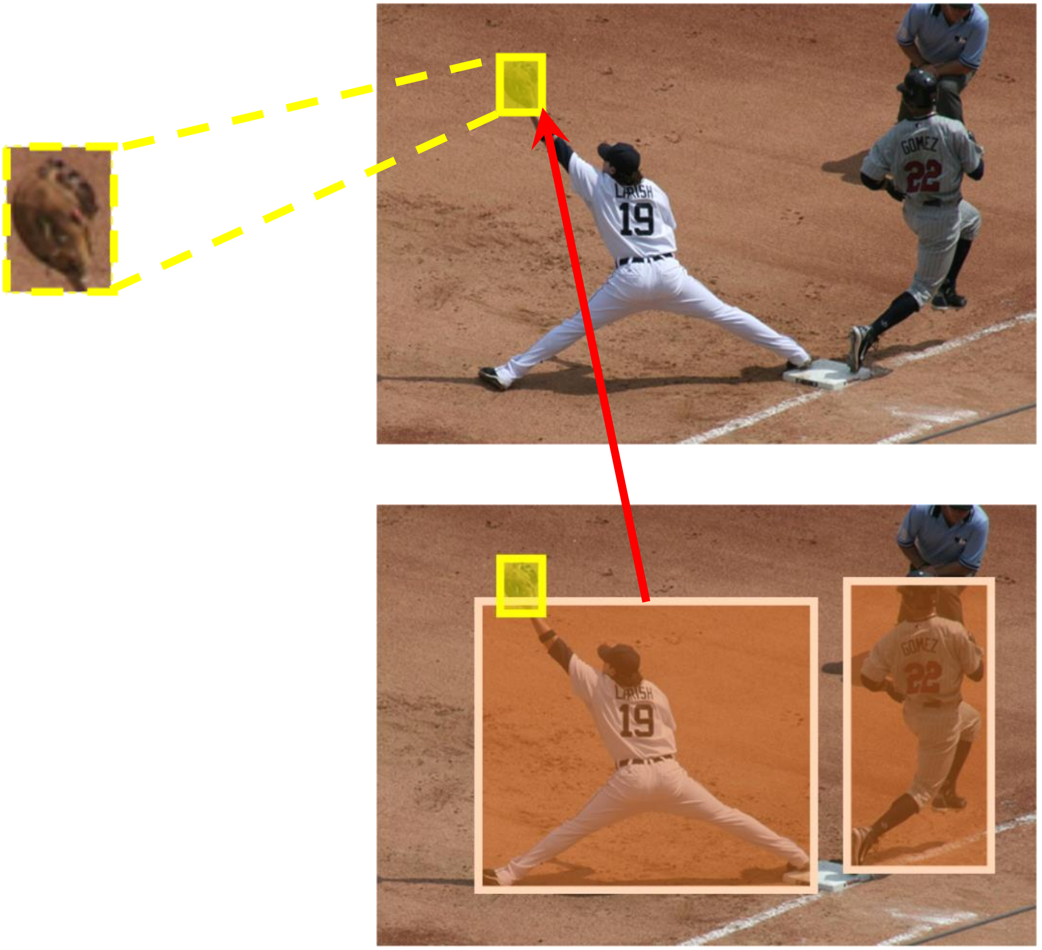
Relation Networks for **Visual Modeling**



Object-Object Relation Modeling

Han Hu*, Jiayuan Gu*, Zheng Zhang*, Jifeng Dai and Yichen Wei. *Relation Networks for Object Detection.* CVPR 2018

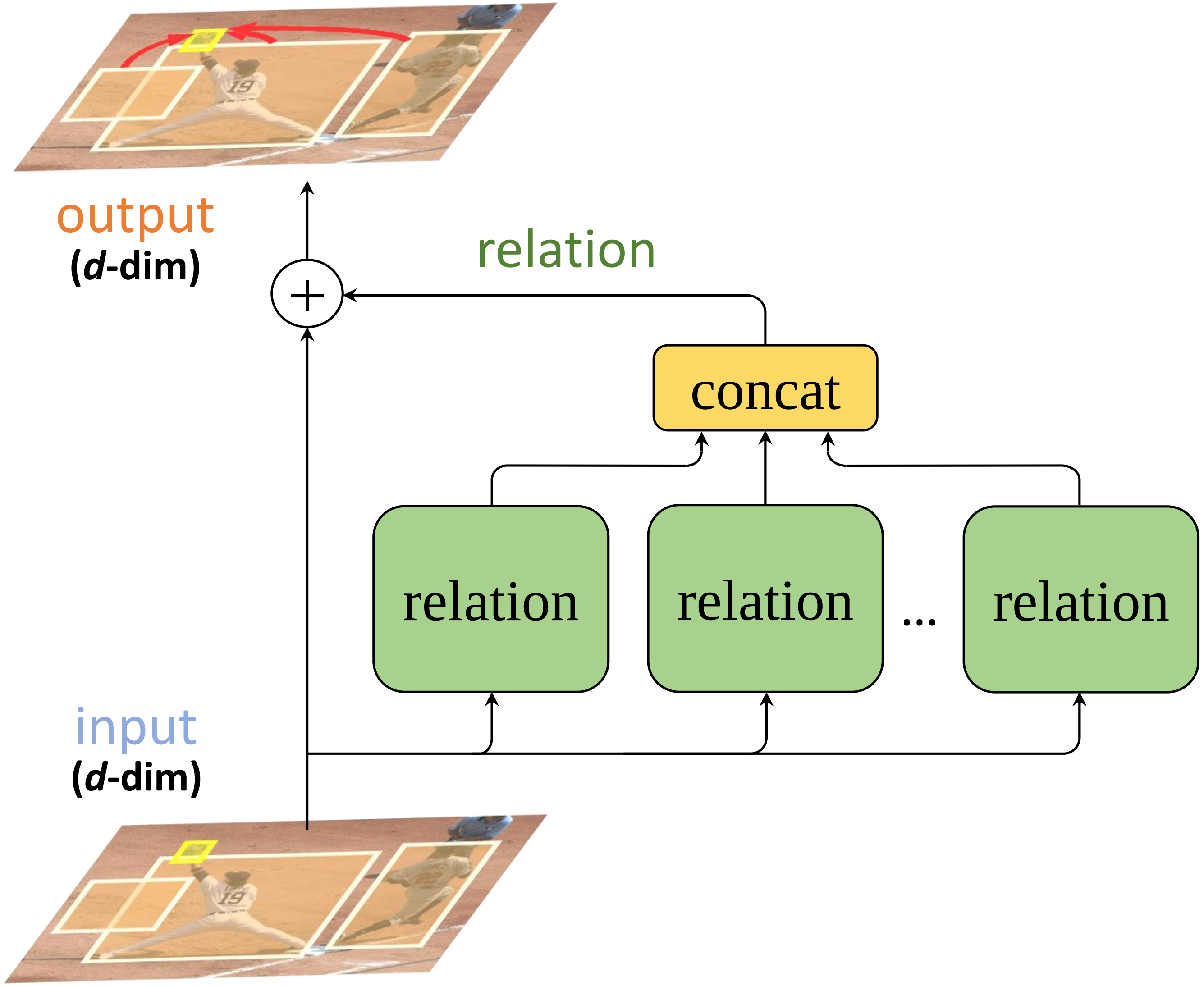
Object-Object Relation Modeling



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

Han Hu*, Jiayuan Gu*, Zheng Zhang*, Jifeng Dai and Yichen Wei. *Relation Networks for Object Detection.* CVPR 2018

Object Relation Module



Plug-and-Play

✓ Parallel, learnable, no additional supervision, translational invariant, stackable

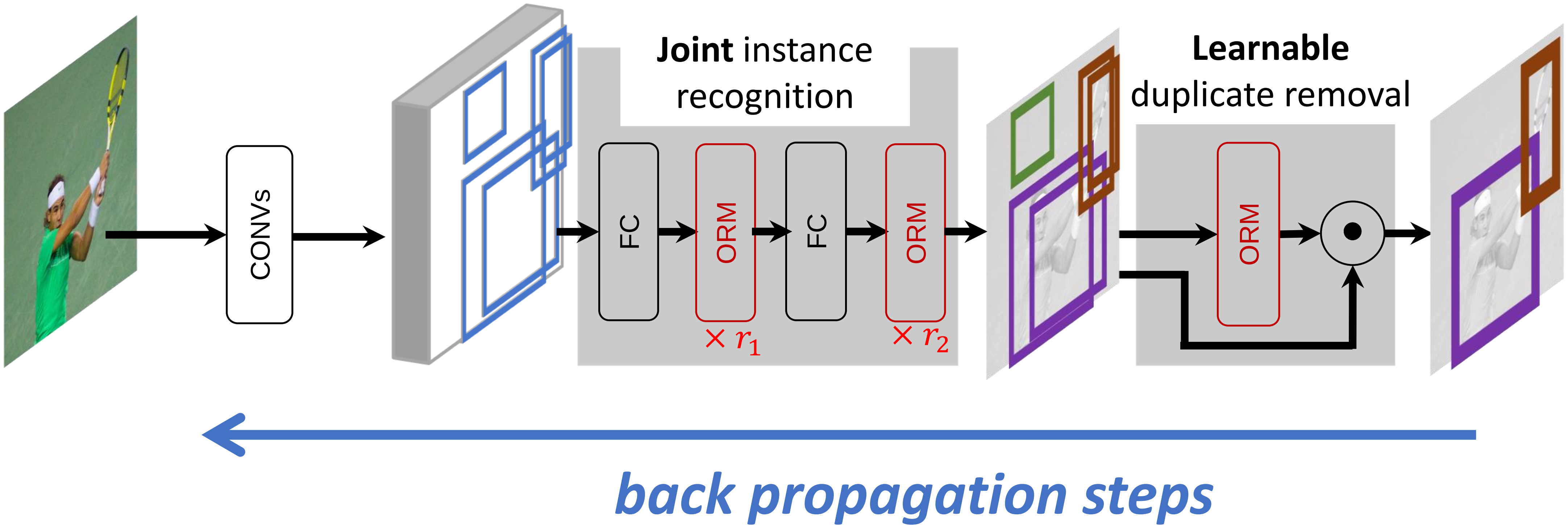
Key Feature

✓ **Relative Geometric** Term

✓ Multiple Relation Branches

✓ Shortcut

The First Fully End-to-End Object Detector



S. Ren et al. *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.* NIPS 2015

Results on COCO Object Detection

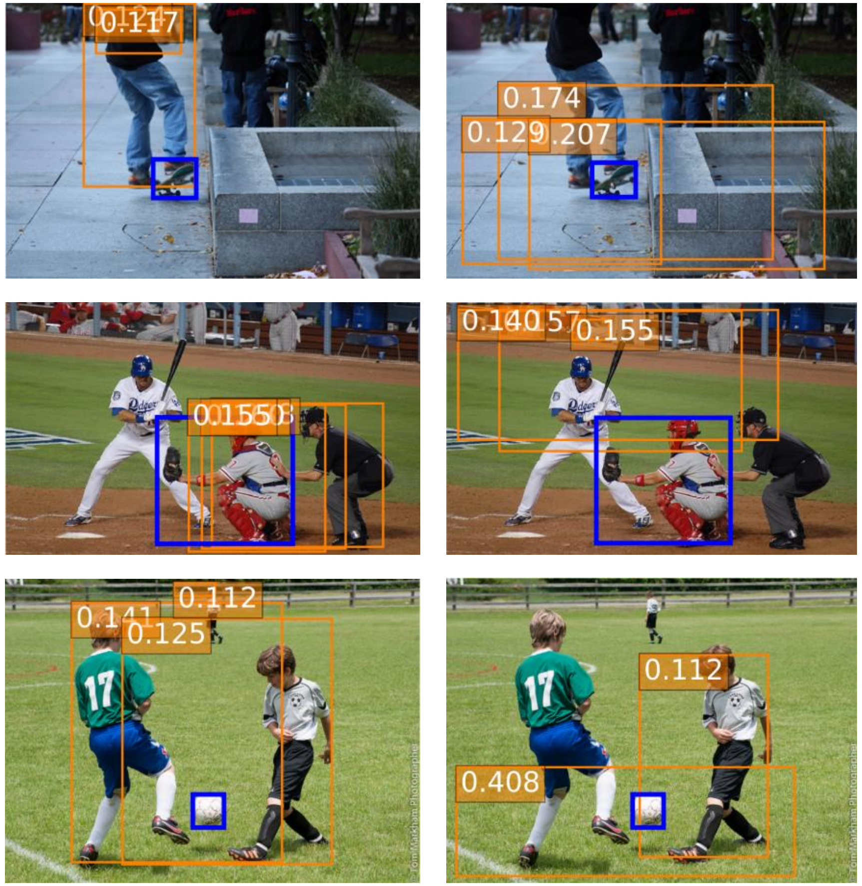
backbone	setting	mAP	mAP_{50}	mAP_{75}	#. params	FLOPS	
faster RCNN	2fc+SoftNMS	32.2/32.7	52.9/53.6	34.2/34.7	58.3M	122.2B	
	2fc+RM+SoftNMS	34.7/35.2	55.3/**56.2**	37.2/37.8	64.3M	124.6B	+3.0 mAP
	2fc+RM+e2e	**35.2**/**35.4**	**55.8**/56.1	**38.2**/**38.5**	64.6M	124.9B	
FPN	2fc+SoftNMS	36.8/37.2	57.8/58.2	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)

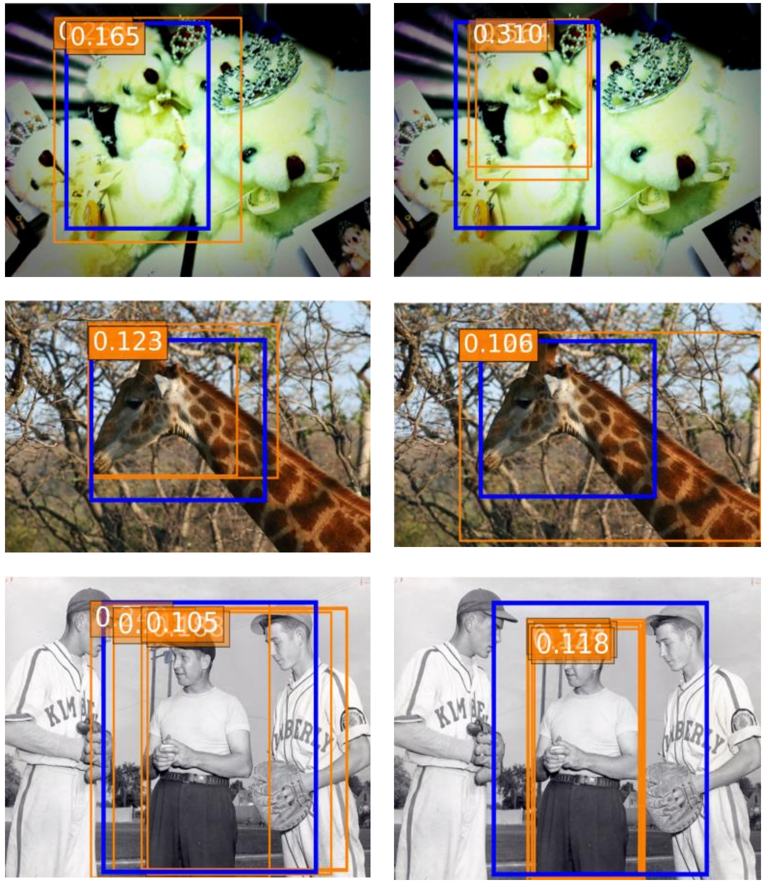
- less than 10% computation overhead on all backbones

Object Pairs with High Relation Weights

instance recognition

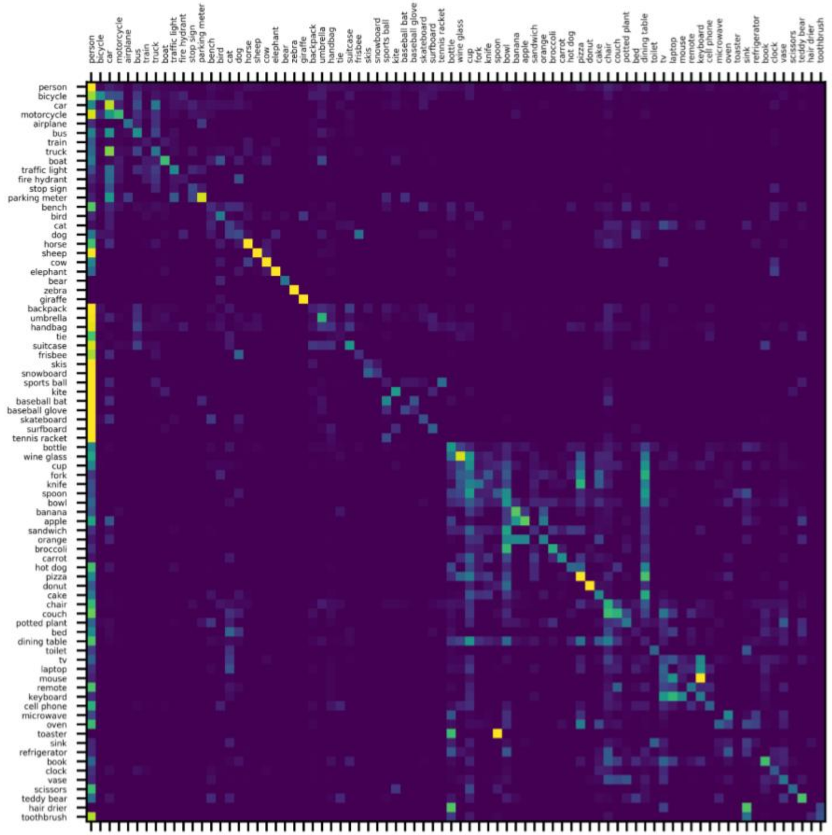


duplicate removal

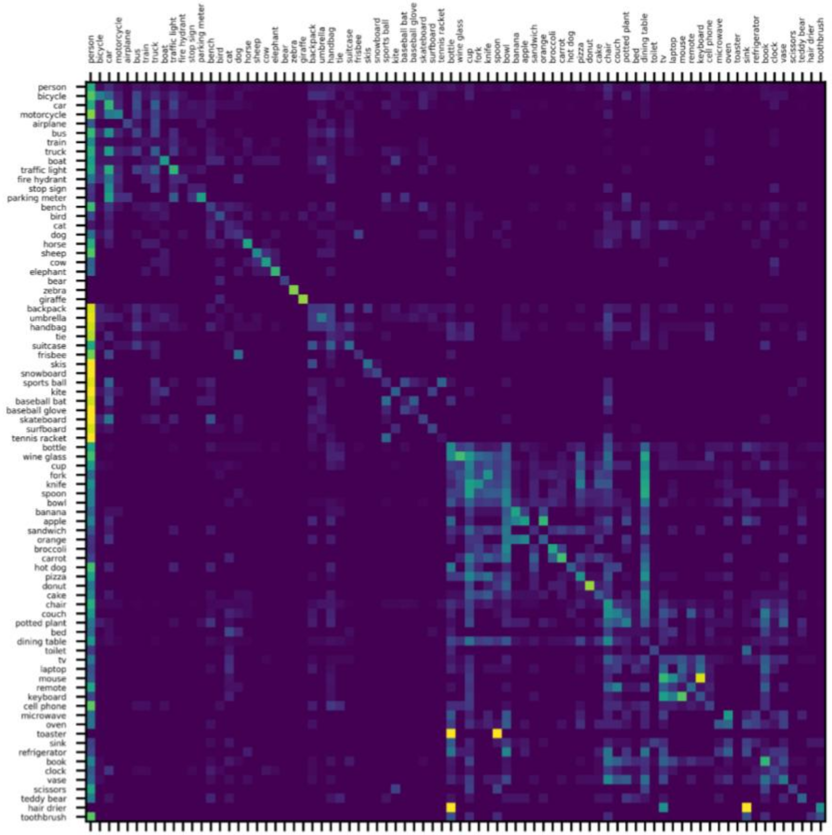


reference object other objects contributing high weights

Class Co-Occurrence Information is Learnt



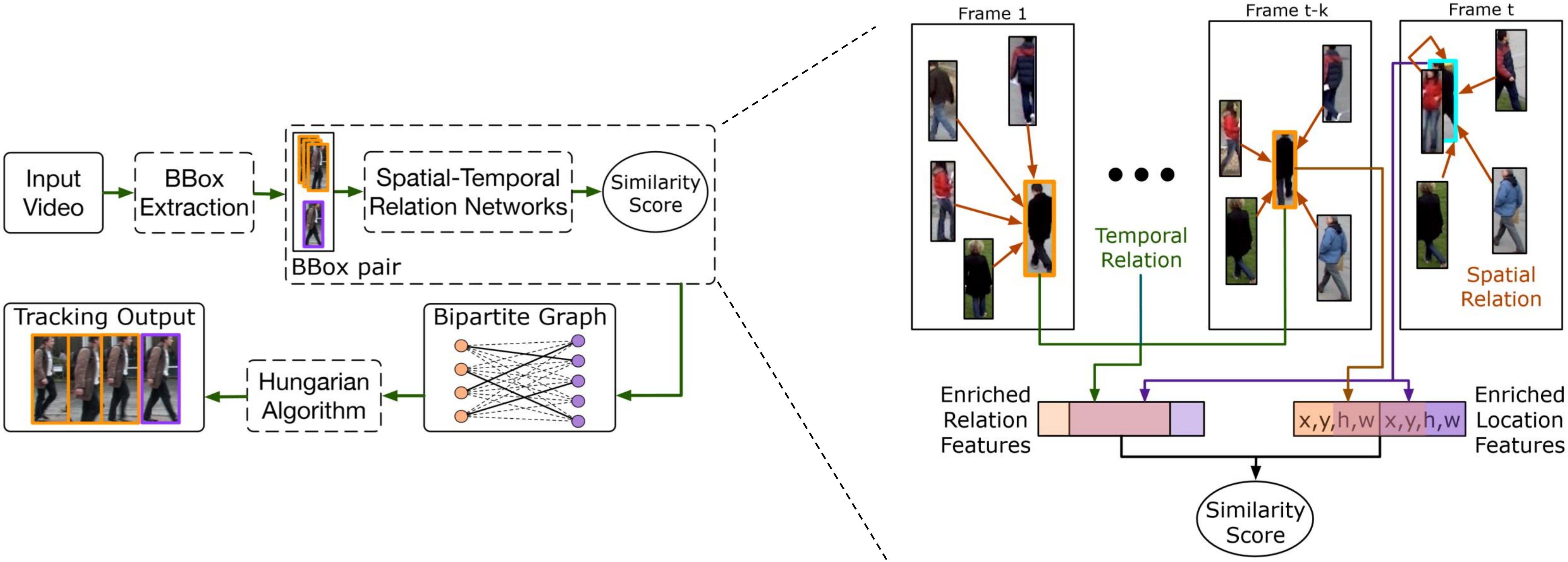
Class Co-occurrence Probability



Learnt Attentional Weights

$r = 0.90$

Extension: Spatial-Temporal Object Relation



Jiarui Xu, Yue Cao, Zheng Zhang and Han Hu. *Spatial-Temporal Relation Networks for Multi-Object Tracking.* Tech Report 2018

Learnable Object-Pixel Relation (vs. RoIAlign)

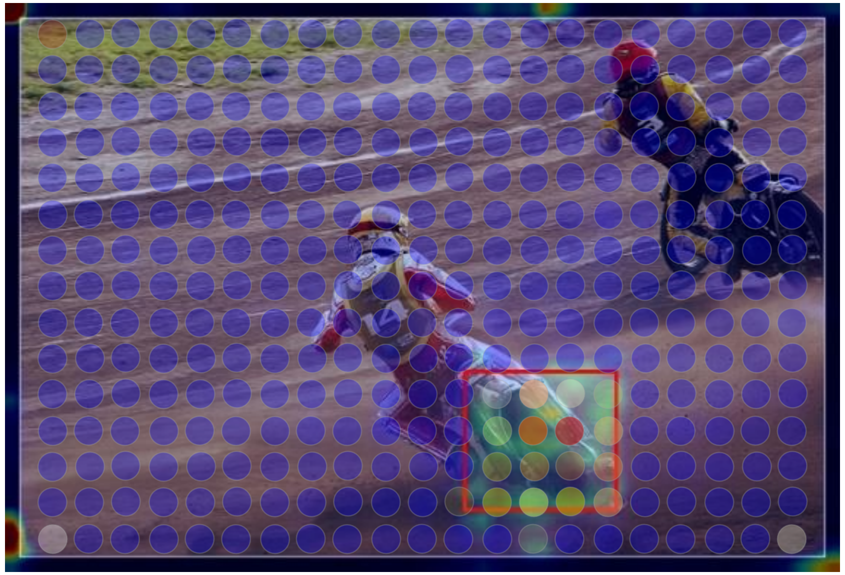


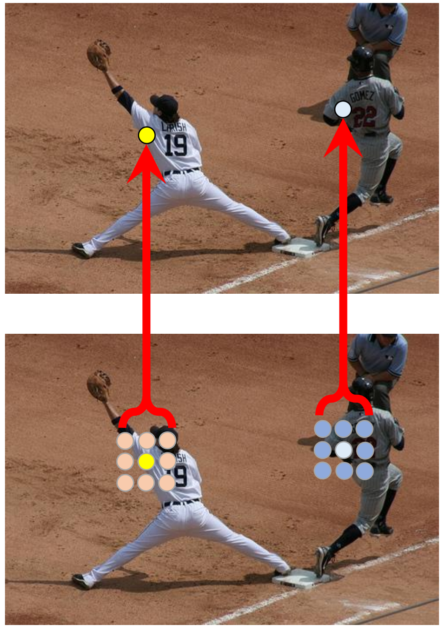
Image Feature to Region Feature

Geometric

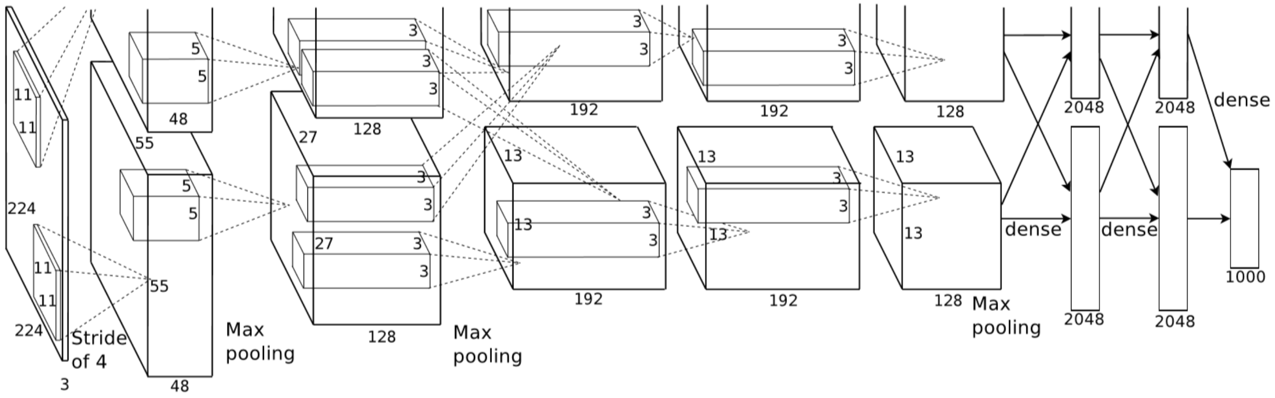
Appearance

Jiayuan Gu, Han Hu, Liwei Wang, Yichen Wei and Jifeng Dai. *Learning Region Features for Object Detection.* ECCV 2018

Pixel-Pixel Relation Modeling



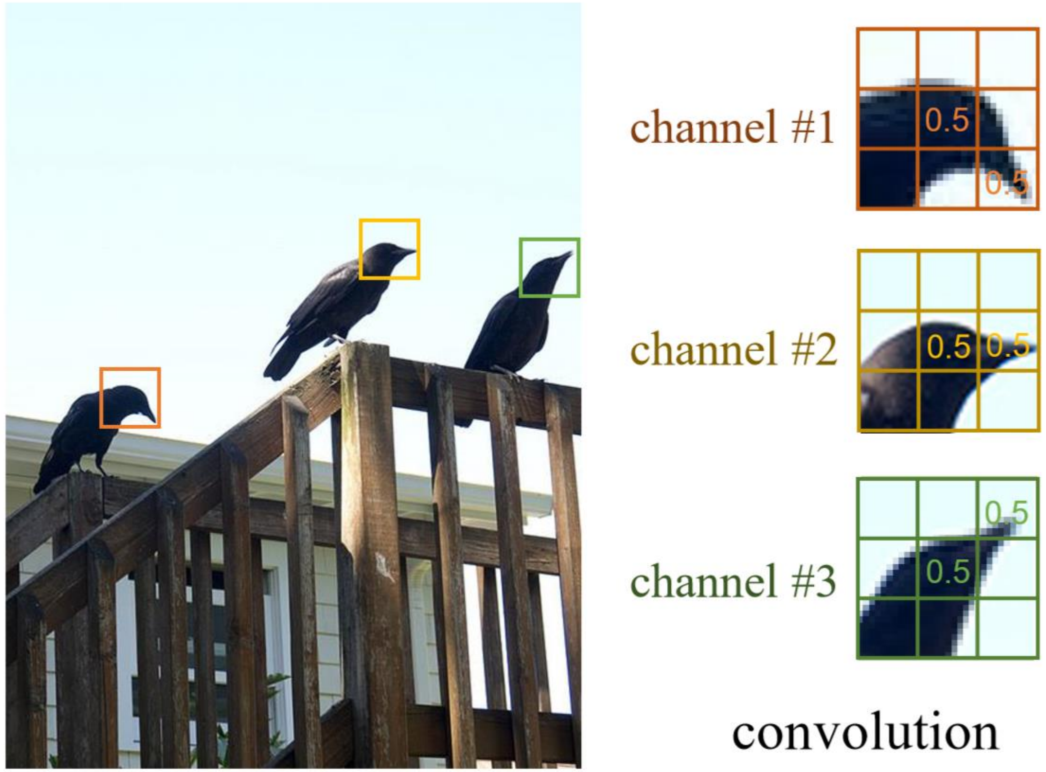
convolution



ConvNets

Question I: Can We Go Beyond *Convolution*?

convolution = template matching



Can we model the patterns by **one** channel?

template -> compose

Han Hu, Zheng Zhang, Zhenda Xie and Stephen Lin. *Local Relation Networks for Visual Recognition. Tech Report 2019*

Related Works: Capsule Networks

- Not aligned well with modern learning infrastructure

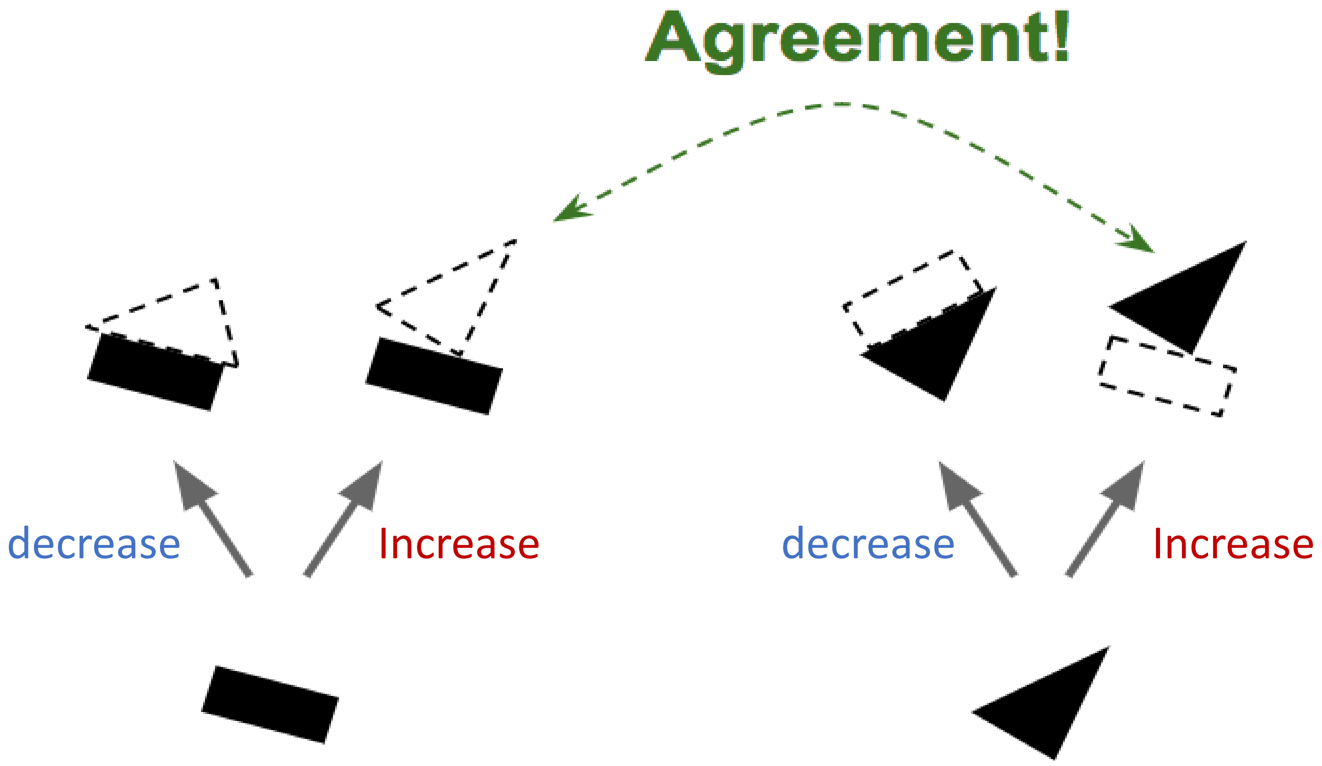
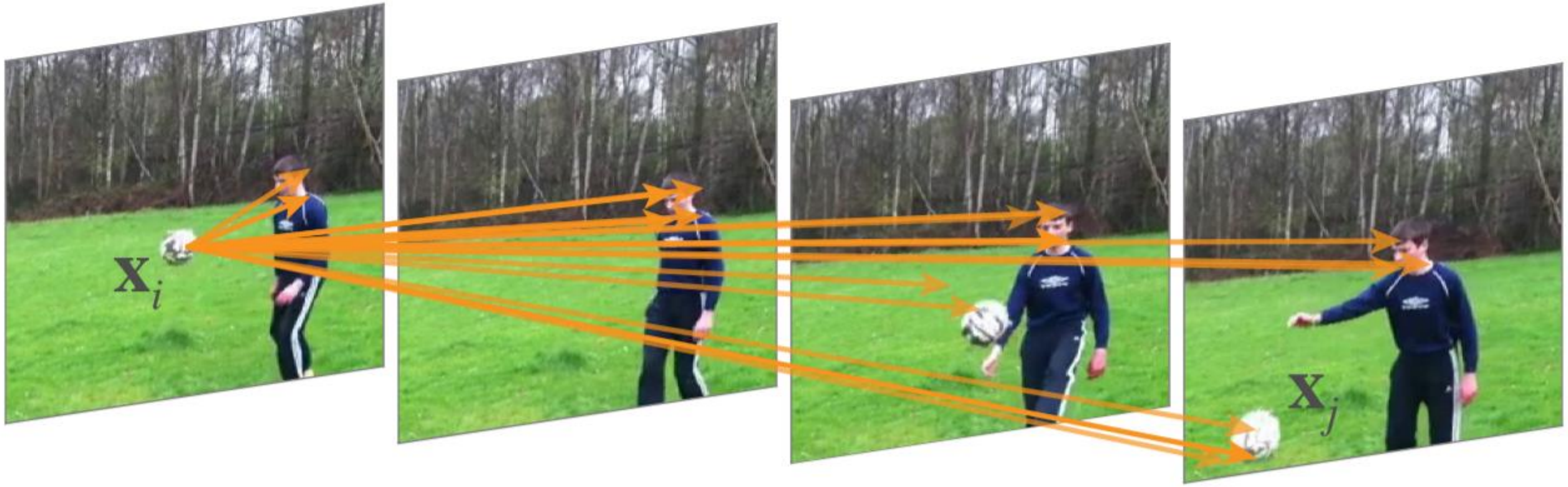


Figure credit by Aurélien Géron

S. Sabour et al. *Dynamic Routing Between Capsules*. NIPS2017

Related Works: Non-Local Neural Networks

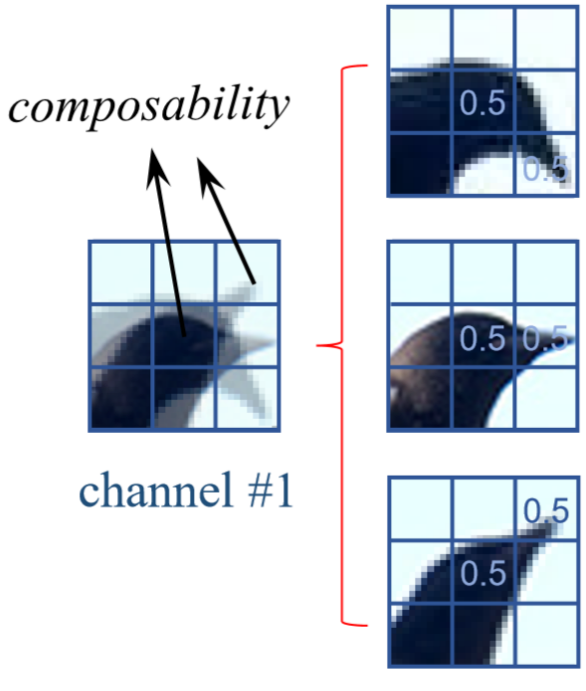
- Complementary to ConvNets



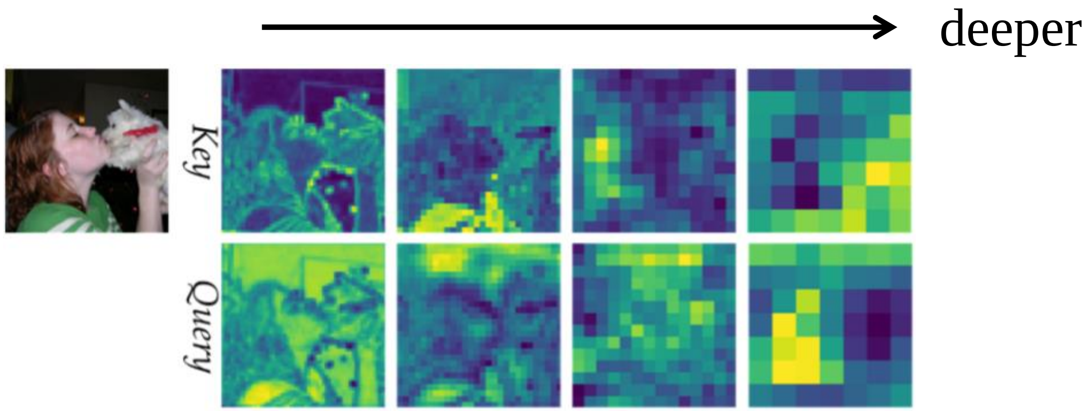
X. Wang et al. *Non-local Neural Networks*. CVPR2018

Beyond Convolution: Local Relation Layer

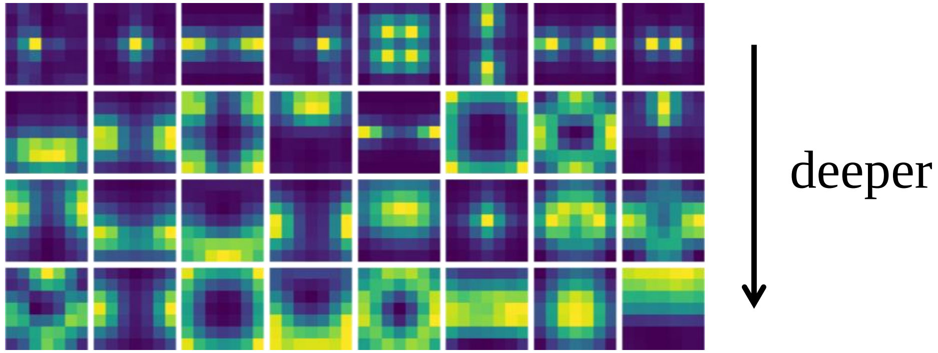
= relation network + locality + geometric prior + scalar key/query



local relation



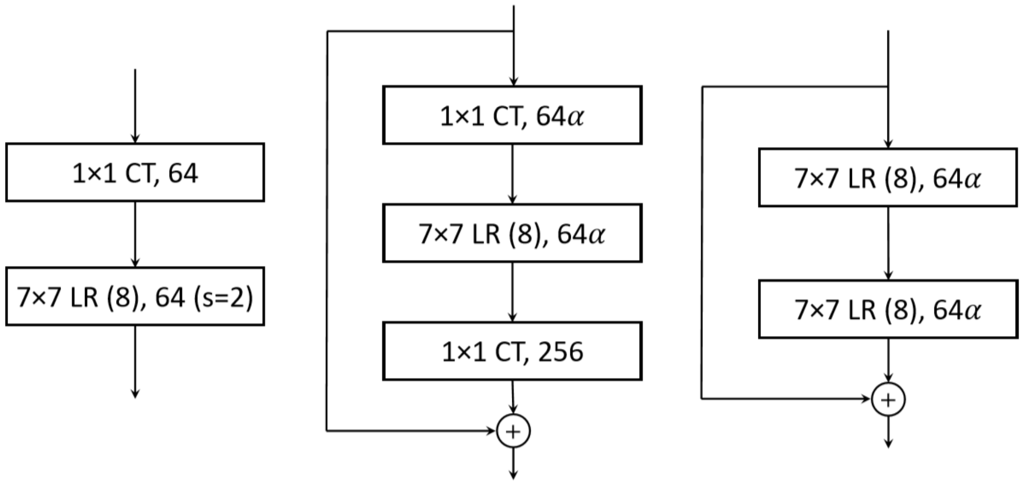
appearance composability



geometric composability (prior)

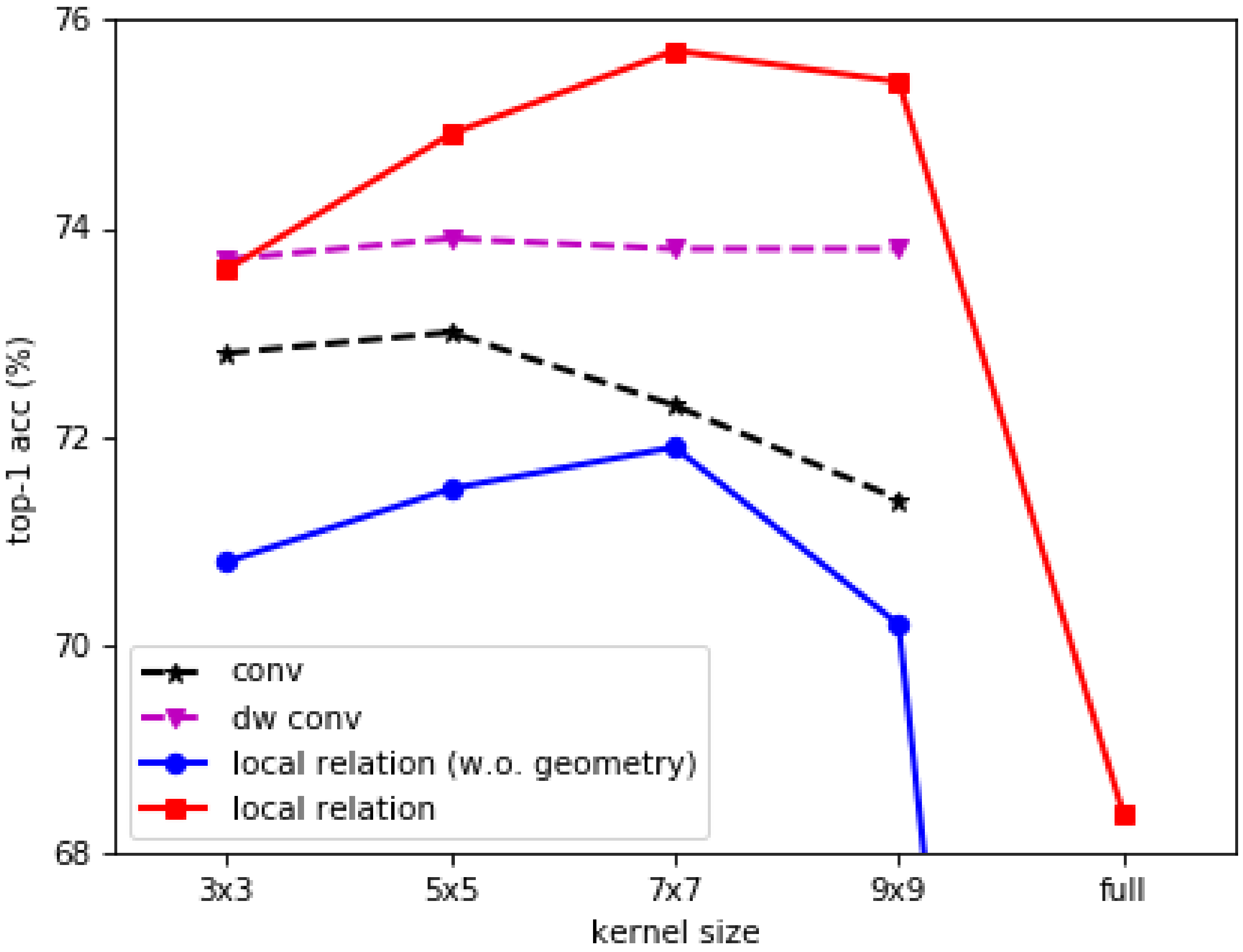
Local Relation Network (LR-Net)

stage	output	ResNet-50	**LR-Net-50 (7×7, *m*=8)**
res1	112×112	7×7 conv, 64, stride 2	**1×1, 64** **7×7 LR, 64, stride 2**
res2	56×56	3×3 max pool, stride 2 [1×1, 64 3×3 conv, 64 1×1, 256] ×3	3×3 max pool, stride 2 [1×1, 100 **7×7 LR, 100** 1×1, 256] ×3
res3	28×28	[1×1, 128 3×3 conv, 128 1×1, 512] ×4	[1×1, 200 **7×7 LR, 200** 1×1, 512] ×4
res4	14×14	[1×1, 256 3×3 conv, 256 1×1, 1024] ×6	[1×1, 400 **7×7 LR, 400** 1×1, 1024] ×6
res5	7×7	[1×1, 512 3×3 conv, 512 1×1, 2048] ×3	[1×1, 800 **7×7 LR, 800** 1×1, 2048] ×3
	1×1	global average pool 1000-d fc, softmax	global average pool 1000-d fc, softmax
# params		$\mathbf{25.5}\times10^6$	$\mathbf{23.3}\times10^6$
FLOPs		$\mathbf{4.3}\times10^9$	$\mathbf{4.3}\times10^9$



Totally convolution free!

Classification on ImageNet (26 Layers)

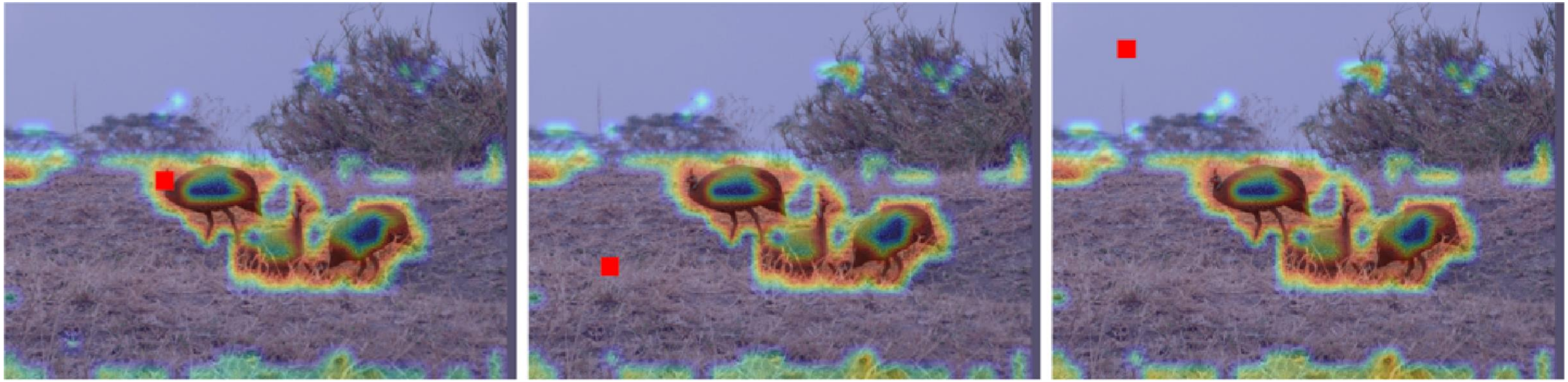


Robust to Adversarial Attacks

network	adversarial train			regular train
	clean	*targeted*	*untargeted*	clean
ResNet-26	44.9	37.9	14.4	72.8
ResNet-50	52.0	43.0	22.5	**76.3**
LR-Net-26	**52.1**	**44.2**	**26.8**	75.7

Question II: Do Non-local Networks Work Well Due to Relation Learning?

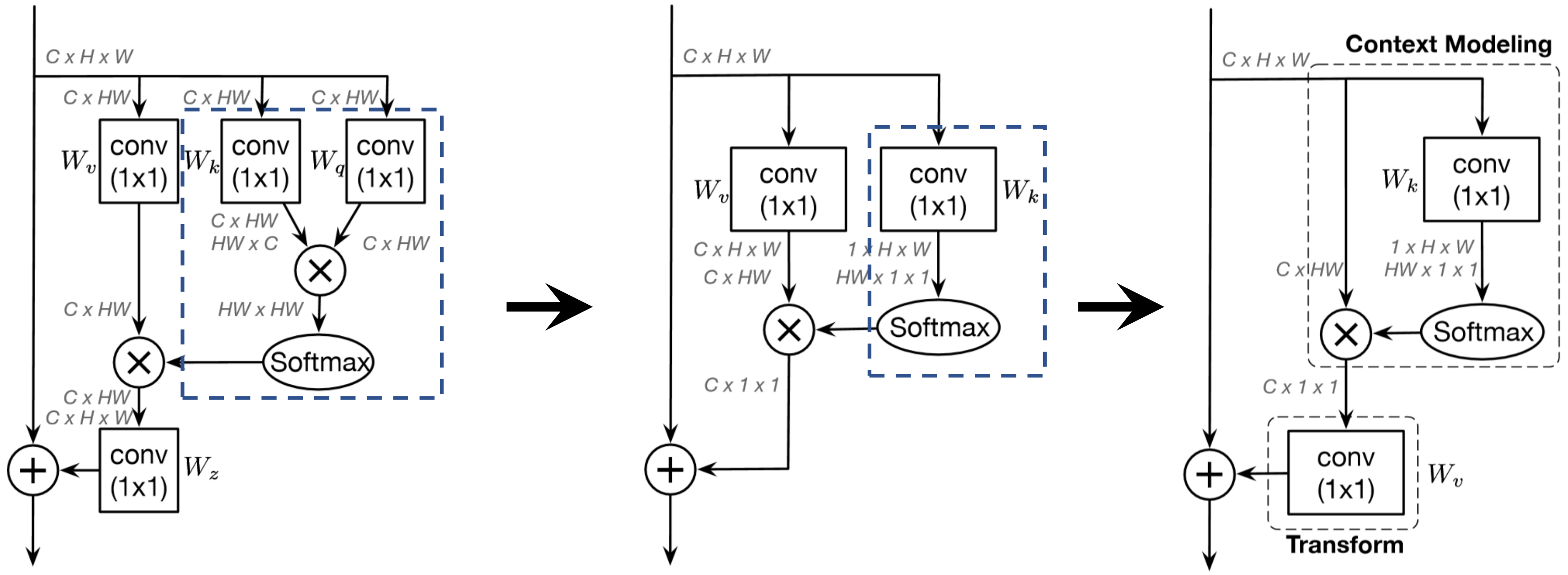
attention maps for different query pixels



Yue Cao*, Jiarui Xu*, Stephen Lin, Fangyun Wei and **Han Hu**.
GCNet: Non-local Networks meet SE-Net and Beyond. Tech Report 2019

Explicit Query-Independent Attention Map

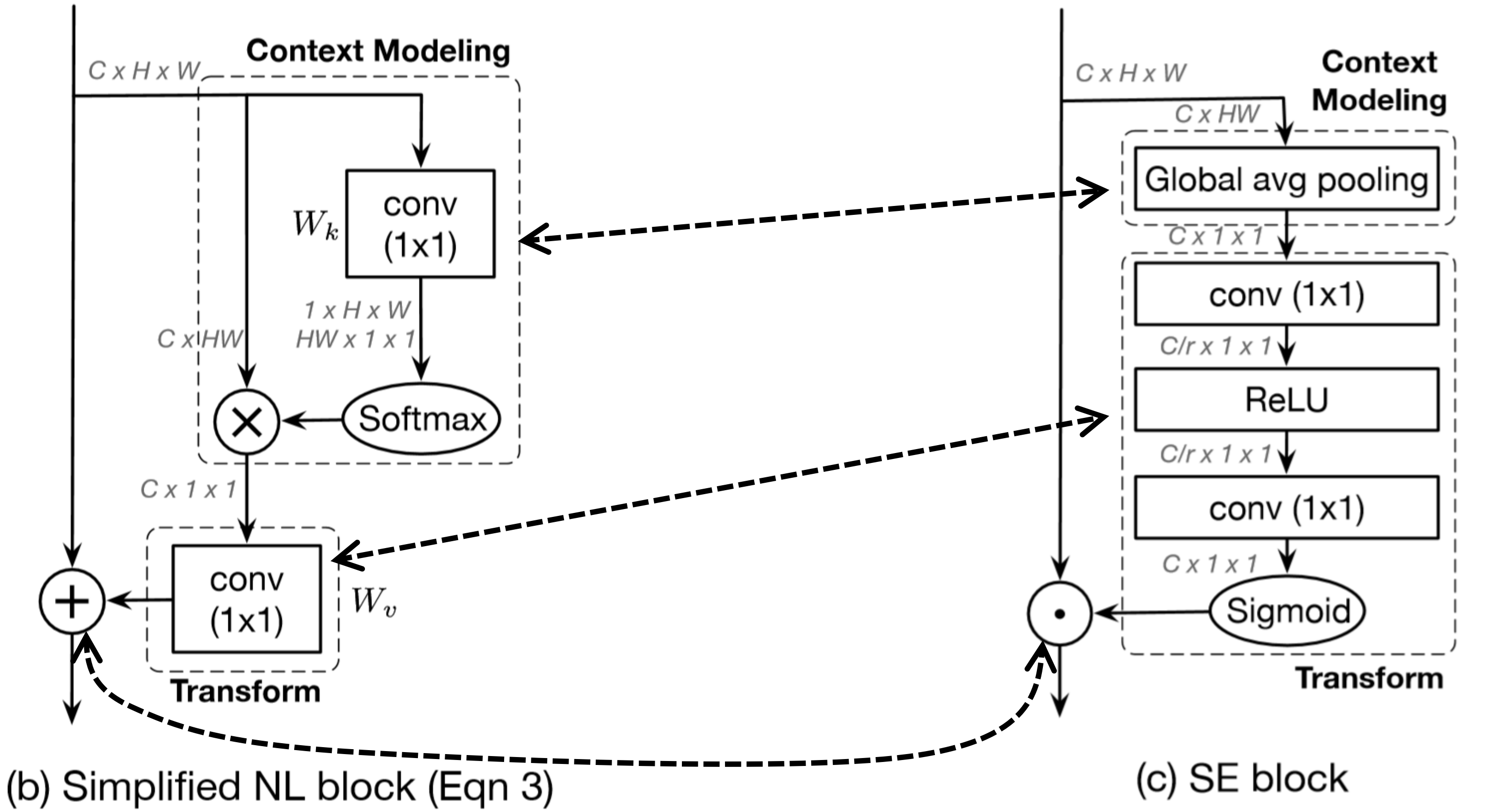
- Simplified Non-Local Blocks



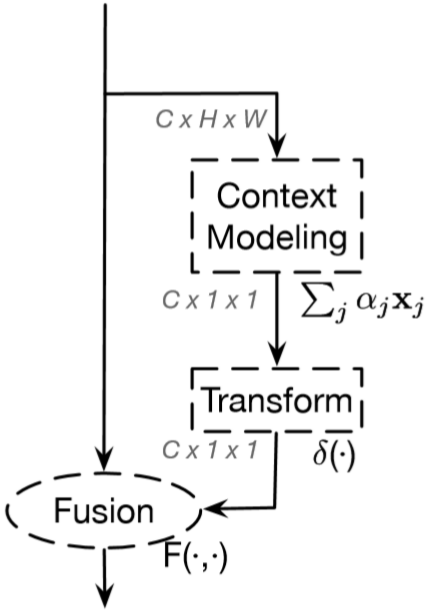
(a) NL block (b) Simplified NL block (Eqn 2) (b) Simplified NL block (Eqn 3)

The same accurate but significantly reducing computation!

Meet SE-Net (2017 ImageNet Champion)



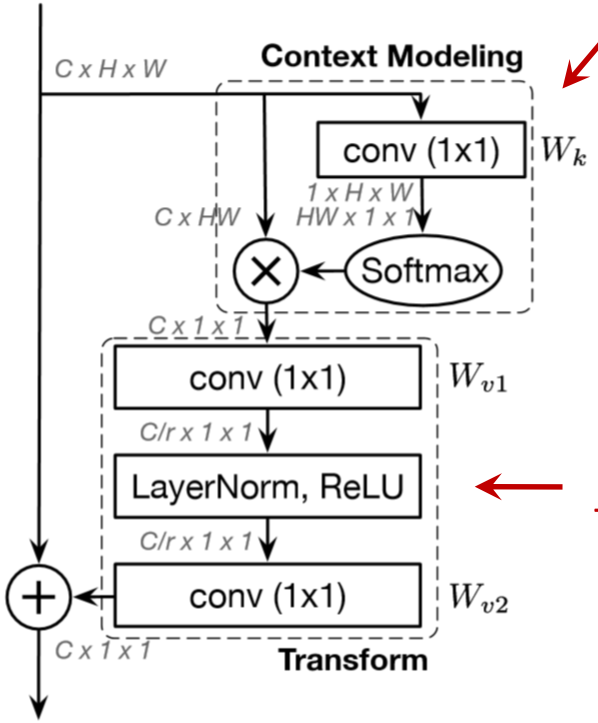
Abstraction and New Instantiation



(a) Global context modeling framework

new instantiation →

1. Global Attention Pooling (Simplified NL-Net)



(d) Global context (GC) block

2. Bottleneck transform (SE-Net)

3. Addition (Simplified NL-Net)

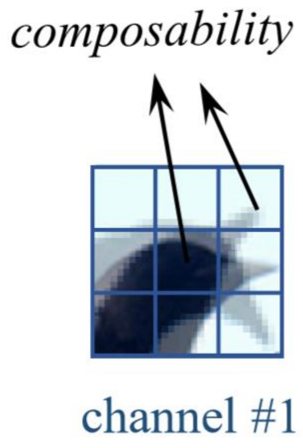
COCO Object Detection Results

- Baseline: Mask R-CNN + ResNet50 + FPN

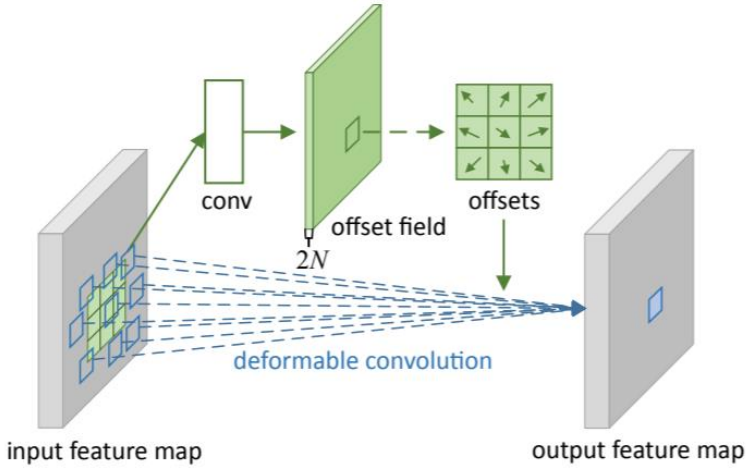
method	AP (bbox)	AP (mask)	#param	FLOPs
baseline	37.2	33.8	44.4M	279.4G
NL-Net	38.0	34.7	46.5M	288.7G
SE-Net	38.2	34.7	46.9M	279.5G
GC-Net	**39.4**	**35.7**	46.9M	279.6G

Discussion: versus Deformable ConvNets

- Both can model content aware adaptiveness
- Verification vs. Regression
- Generality (arbitrary vs. grid)
- Partly complementary



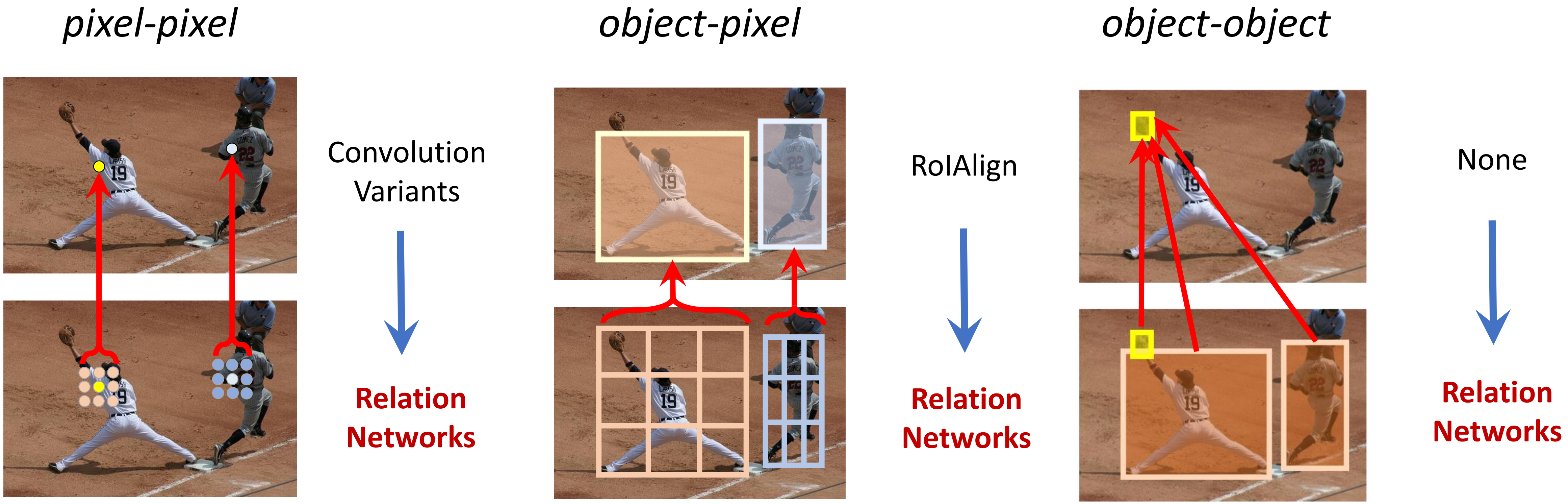
relation networks



deformable conv

[1] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu and Yichen Wei. Deformable Convolutional Networks. In ICCV 2017.
[2] Xizhou Zhu, Han Hu, Stephen Lin and Jifeng Dai. Deformable ConvNets v2: More Deformable, Better Results. In CVPR 2019.
[3] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang and Stephen Lin. RepPoints: Point Set Representation for Object Detection. Tech Report.

Thanks!



*Relation Network is All You Need for AI——**SkyNet***