Recent Advances on Object Detection in MSRA

Jifeng Dai, Han Hu, Lu Yuan and Yichen Wei

Visual Computing Group, Microsoft Research Asia

Outline

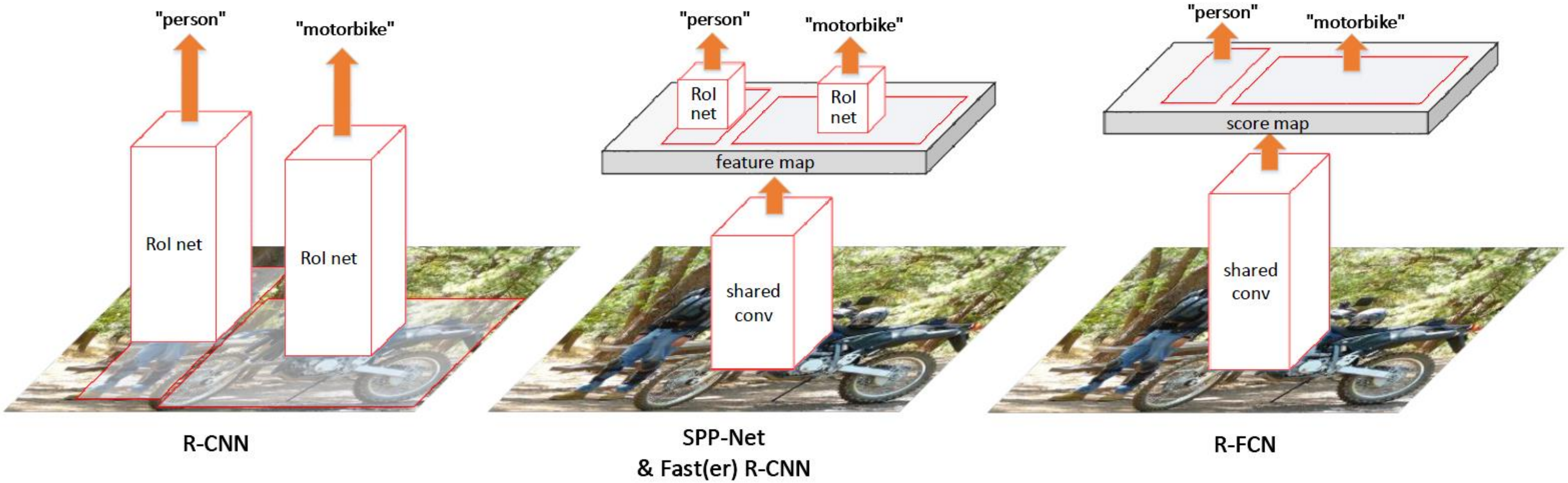
- R-FCN and its extensions
- Deformable ConvNets and its extensions
- Video object detection
- Summary

Highlights

- Region-based, fully-convolutional networks for object detection
- Fast and accurate
- Motivate many extensions

Code is available at https://github.com/daijifeng001/R-FCN

Region-based Object Detectors



- Methodologies of region-based detectors using ResNet-101

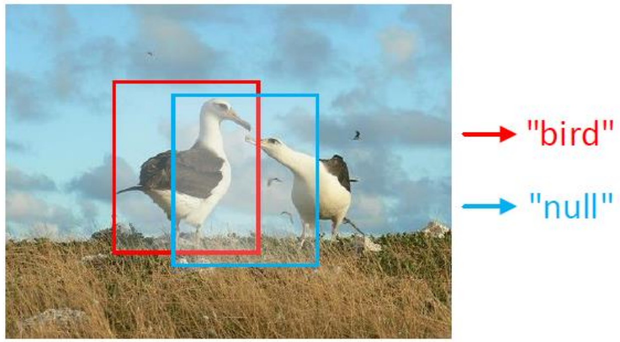
	R-CNN	Faster R-CNN	R-FCN [ours]
depth of shared conv subnetwork	0	91	101
depth of RoI-wise subnetwork	101	10	**0**

Respecting Translation Variance for Detection

- Increasing translation invariance for image classification
  - Shift of an object inside an image should be indiscriminative
  - Leading deep (fully) convolutional architectures are translation-invariant

- Respecting translation variance for object detection
  - Responses should reflect how candidate boxes overlap with objects
  - A considerable deep per-ROI subnet in Faster-RCNN using ResNet-101

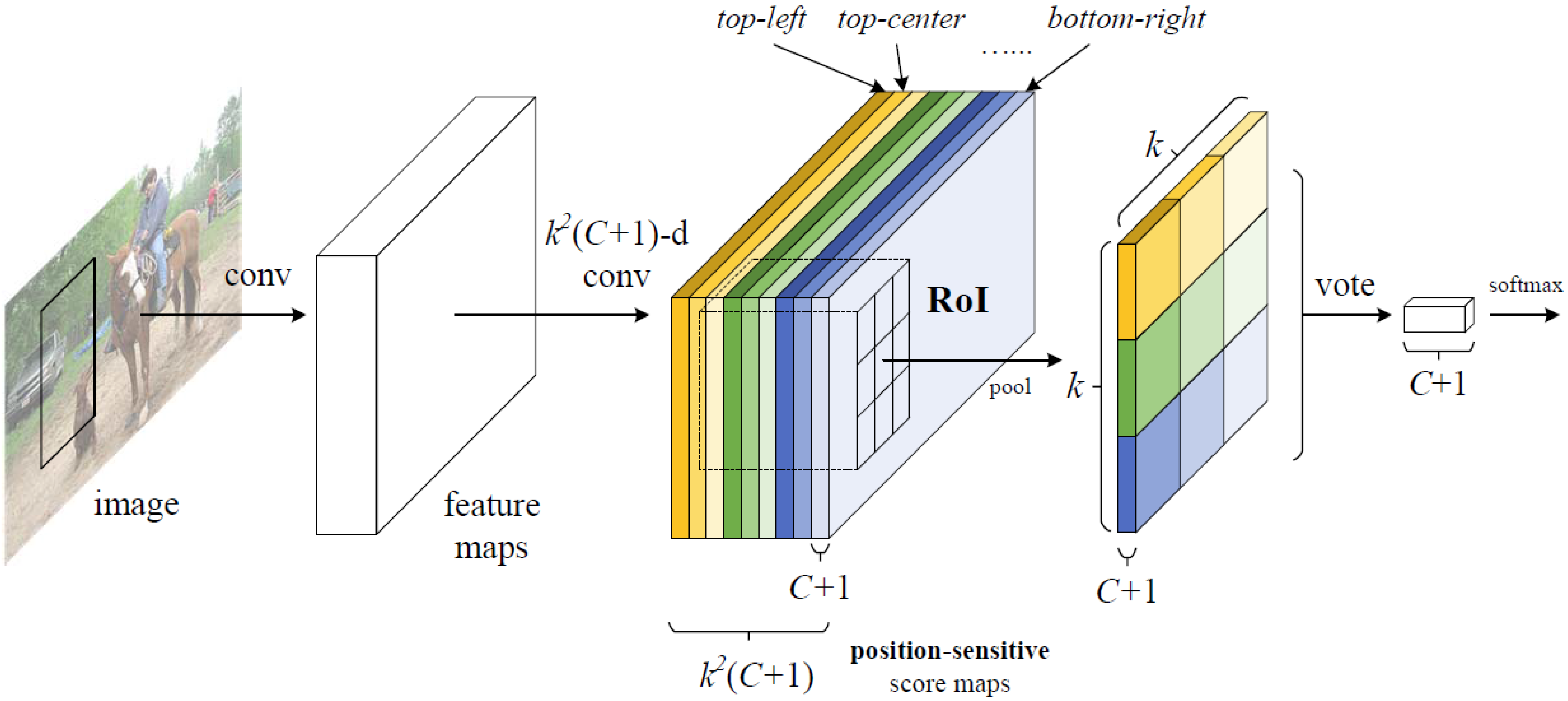
image classification



object detection

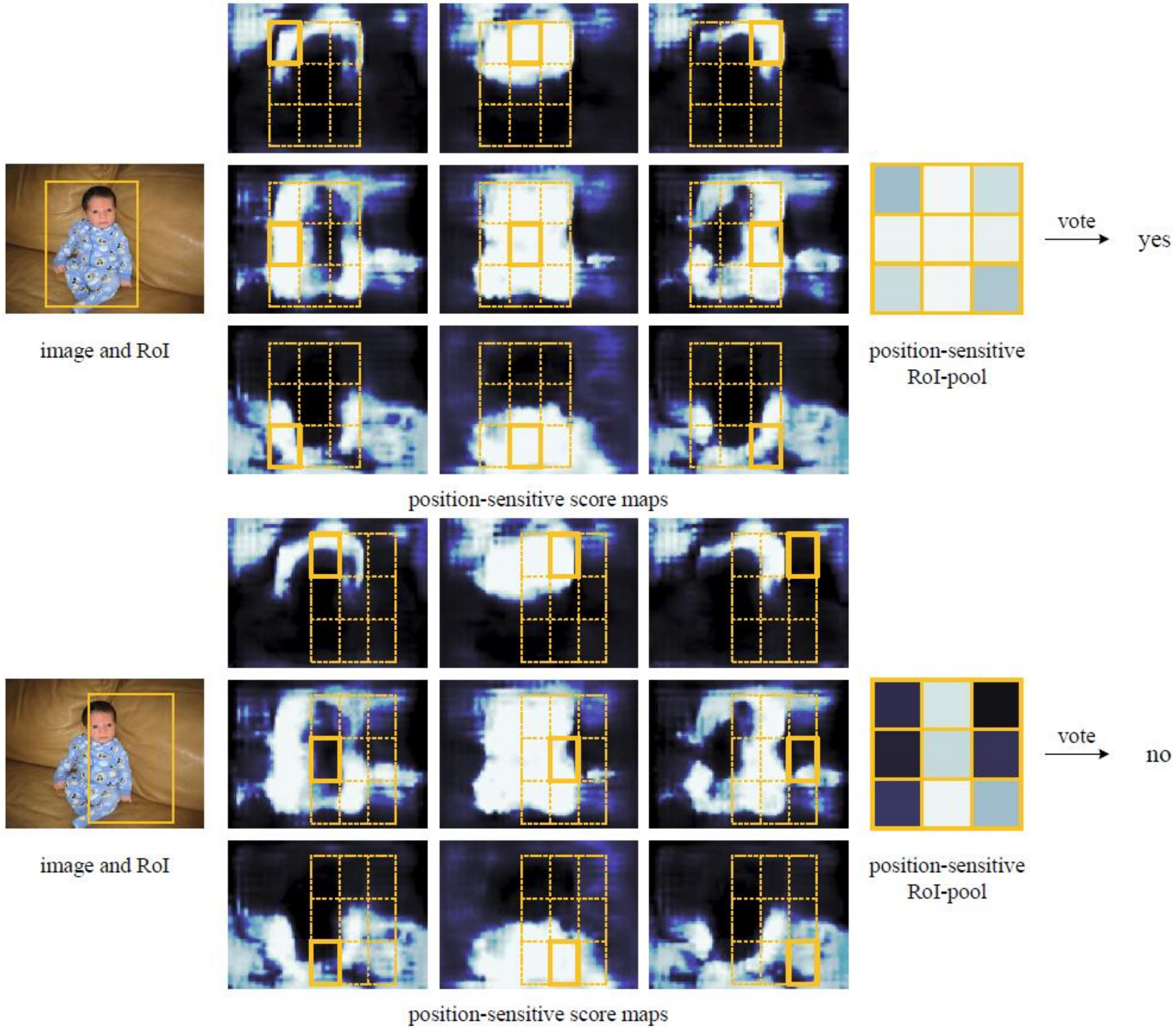
R-FCN

- Key idea of R-FCN for object detection
  - Position-sensitive score maps (kxk, e.g., k = 3)
  - Position-sensitive RoI pooling



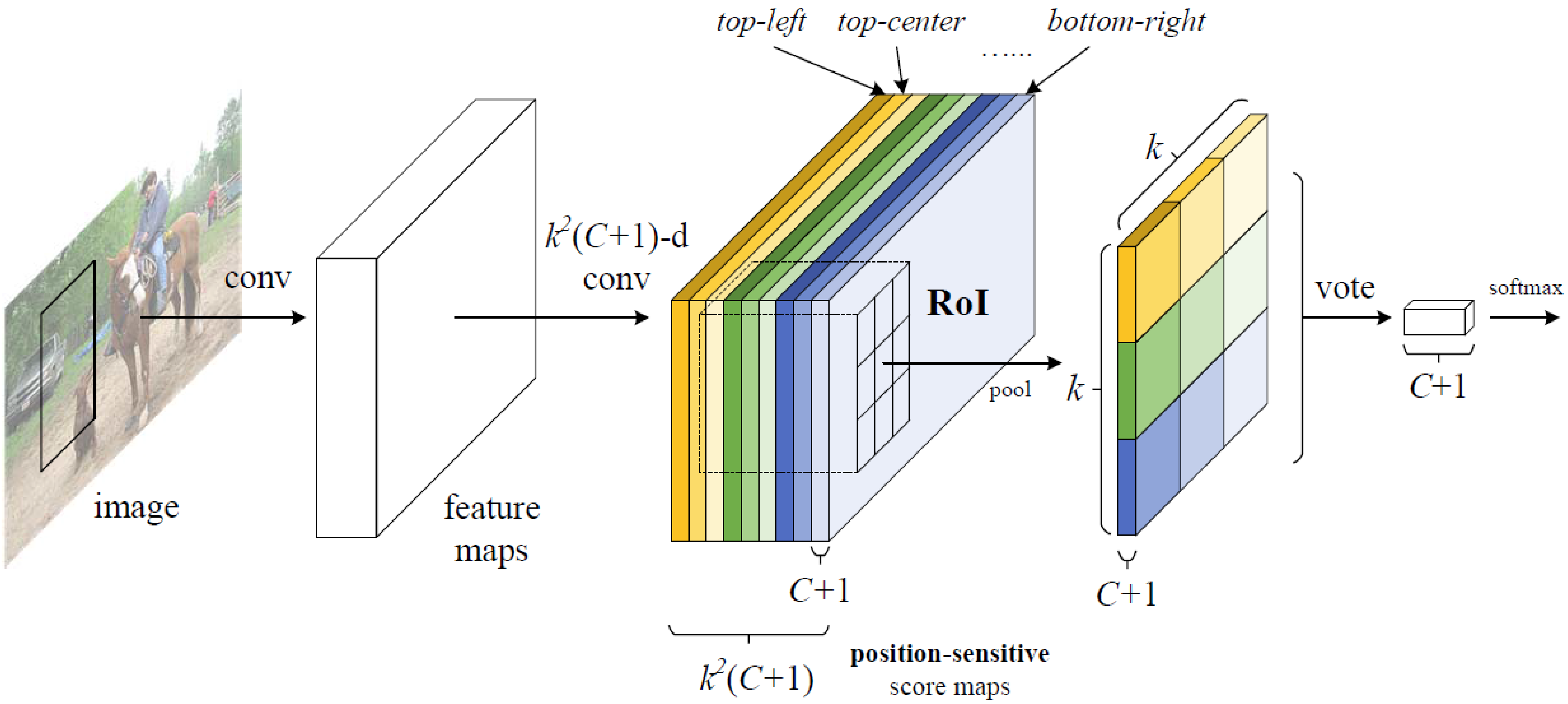
R-FCN

- Spatial information is encoded by position-sensitive score maps



R-FCN

- Key properties of **R-FCN**
  - Negligible per-RoI computational cost (in both training/inference)
  - The whole architecture is end-to-end trainable



Experiments

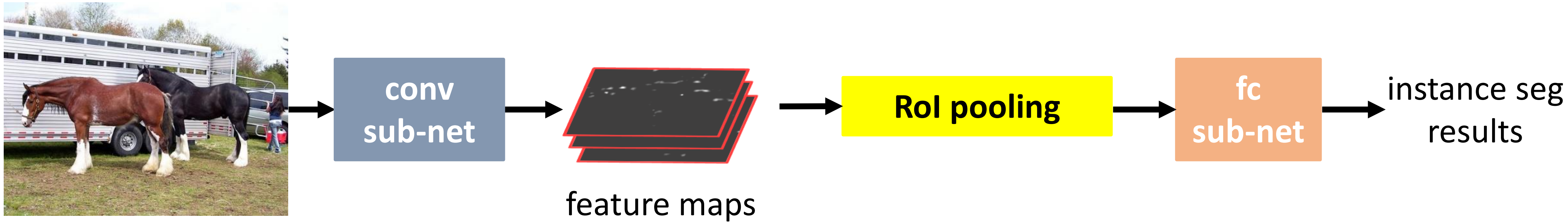
- Comparisons between Faster R-CNN and R-FCN using ResNet-101

	depth of per-RoI subnetwork	training w/ OHEM?	train time (sec/img)	test time (sec/img)	mAP (%) on VOC07
Faster R-CNN	10		1.2	0.42	76.4
R-FCN	0		0.45	0.17	76.6
Faster R-CNN	10	✓(300 RoIs)	1.5	0.42	79.3
R-FCN	0	✓(300 RoIs)	0.45	0.17	**79.5**
Faster R-CNN	10	✓(2000 RoIs)	2.9	0.42	*N/A*
R-FCN	0	✓(2000 RoIs)	0.46	0.17	79.3

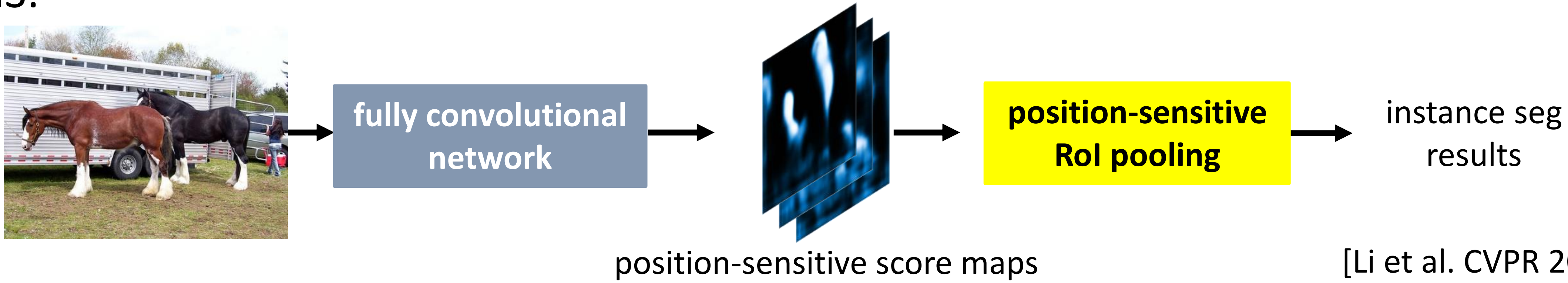
R-FCN extensions: fully convolutional instance segmentation

- First pure fully convolutional solution for instance segmentation
  - Accurate: no feature warping/resizing or fc layers
  - Fast: negligible per-region computation

Previous best & fastest:



FCIS:

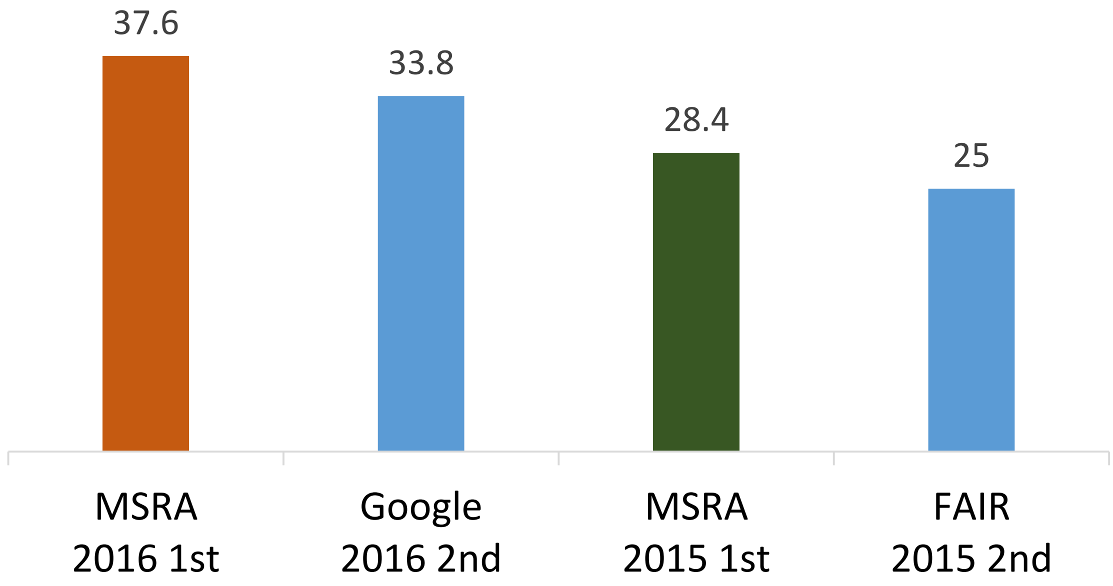


[Li et al. CVPR 2017.]

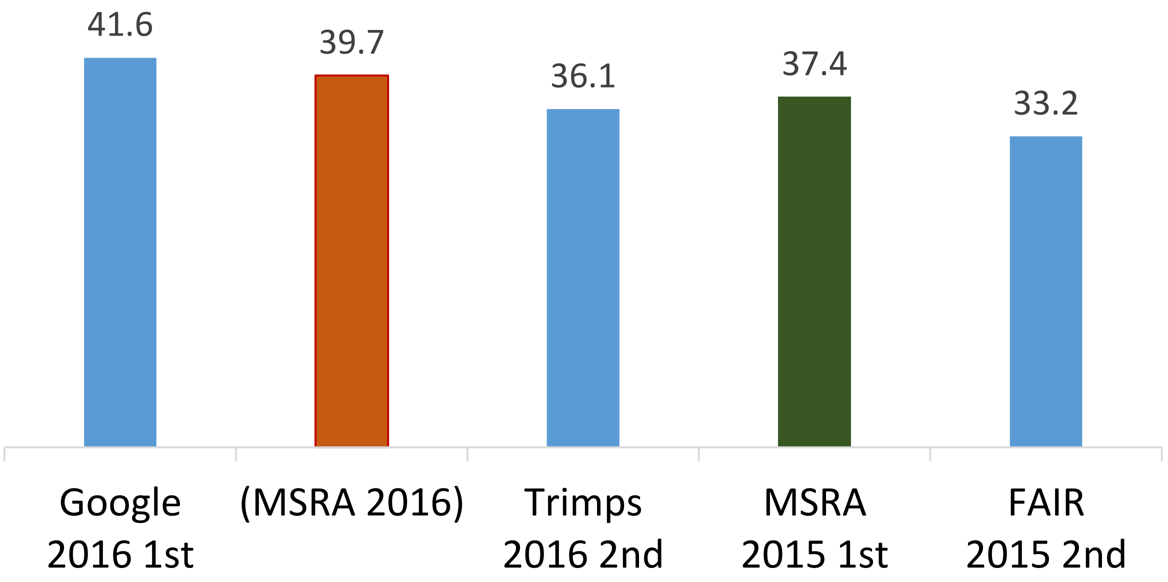
COCO Segmentation Challenge 2016

- MSRA won 1st place back-to-back
  - 11% relatively better than 2016 2nd (Google)
  - 33% relatively better than 2015 1st (MSRA)
  - Excellent on box: 2nd place in detection if public

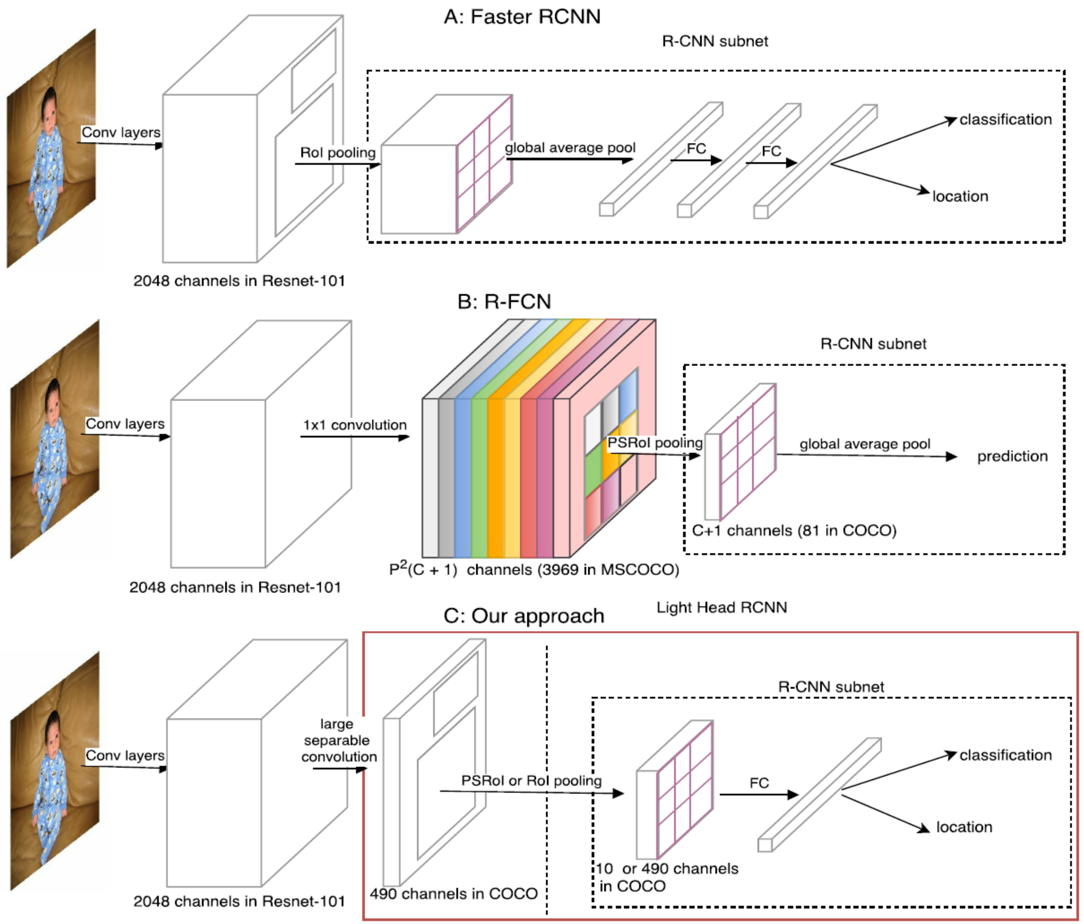
COCO Segmentation Accuracy (%)



COCO Detection Accuracy (%)



R-FCN extensions: Light-head R-CNN

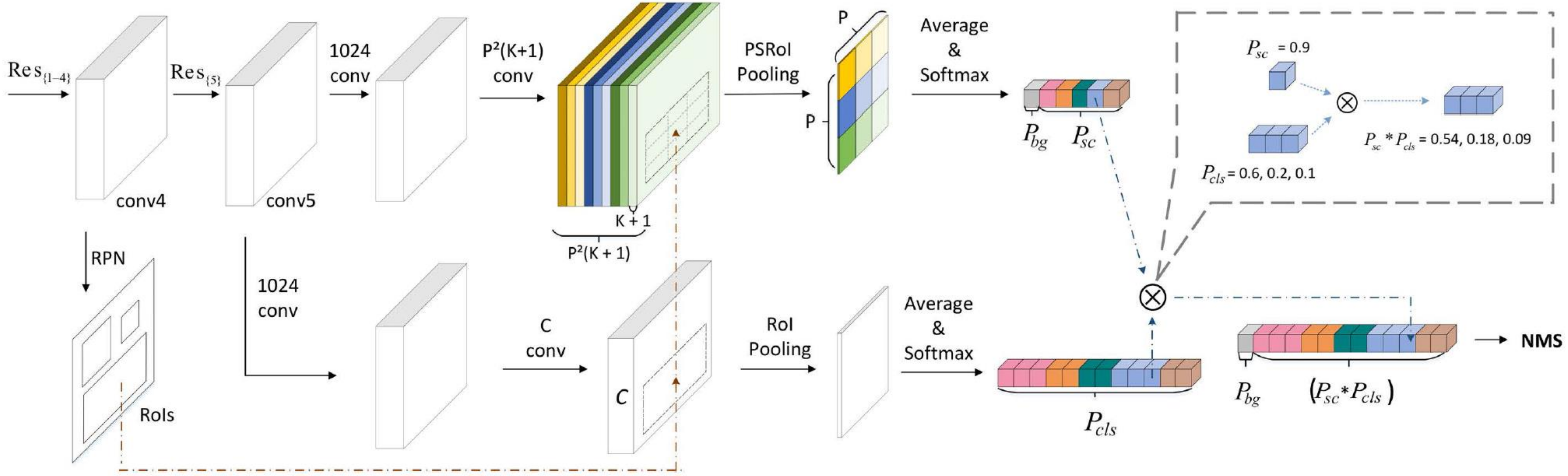


- PS scores-> PS features, followed by ultra-light detection head
  - Fast and accurate
  - Adopted in products

[Li et al. CVPR 2018.]

R-FCN extensions: R-FCN-3000 at 30fps

- Decoupled classification and localization for scaling up



[Singh et al. CVPR 2018.]

Outline

- R-FCN and its extensions
- Deformable ConvNets and its extensions
- Video object detection
- Summary

Highlights

- **Enabling effective modeling of spatial transformation** in ConvNets
- **No additional supervision** for learning spatial transformation
- **Significant accuracy improvements** on sophisticated vision tasks

Code is available at https://github.com/msracver/Deformable-ConvNets

Modeling Spatial Transformations

- A long standing problem in computer vision

Deformation:

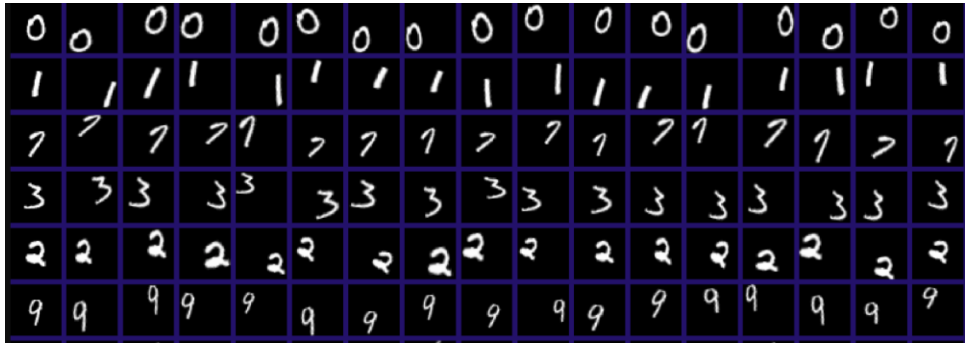
Scale:

Viewpoint variation:

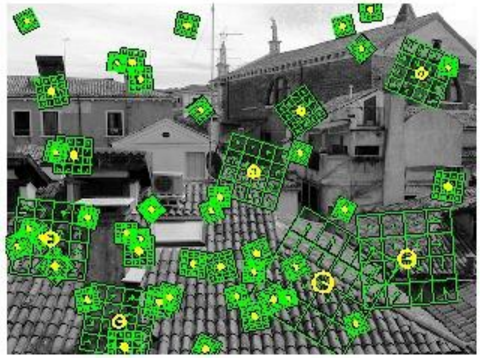
Intra-class variation:

Traditional Approaches

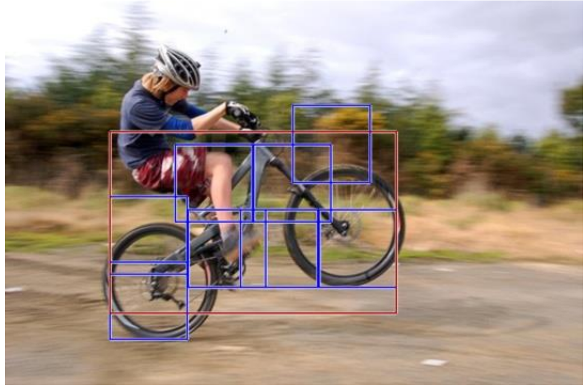
- 1) To build training datasets with sufficient desired variations



- 2) To use transformation-invariant features and algorithms



Scale Invariant Feature Transform (SIFT)

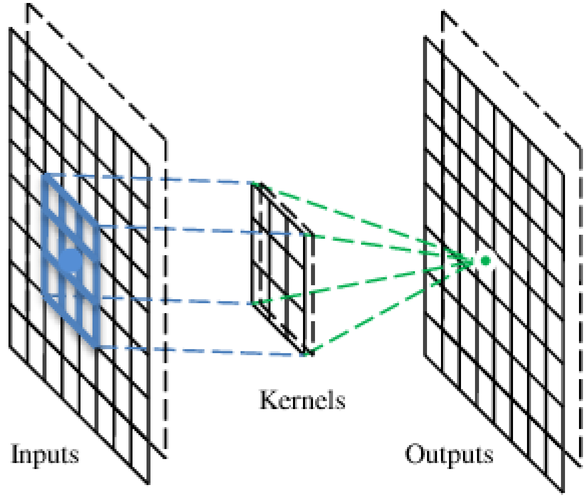


Deformable Part-based Model (DPM)

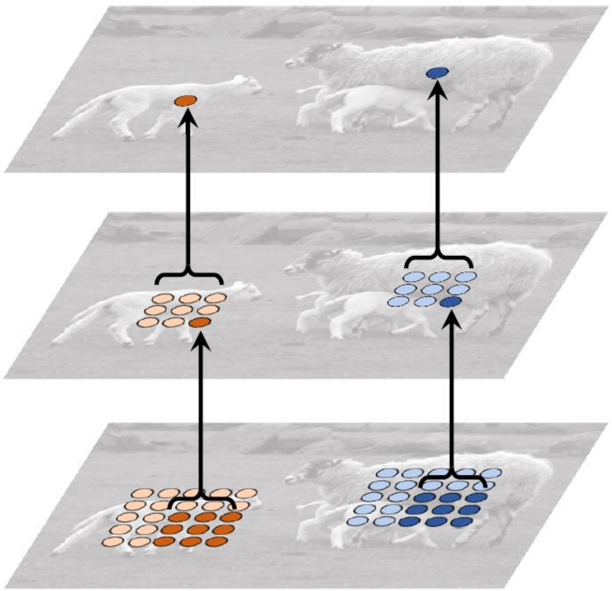
- Drawbacks: geometric transformations are assumed fixed and known, hand-crafted design of invariant features and algorithms

Spatial Transformations in CNNs

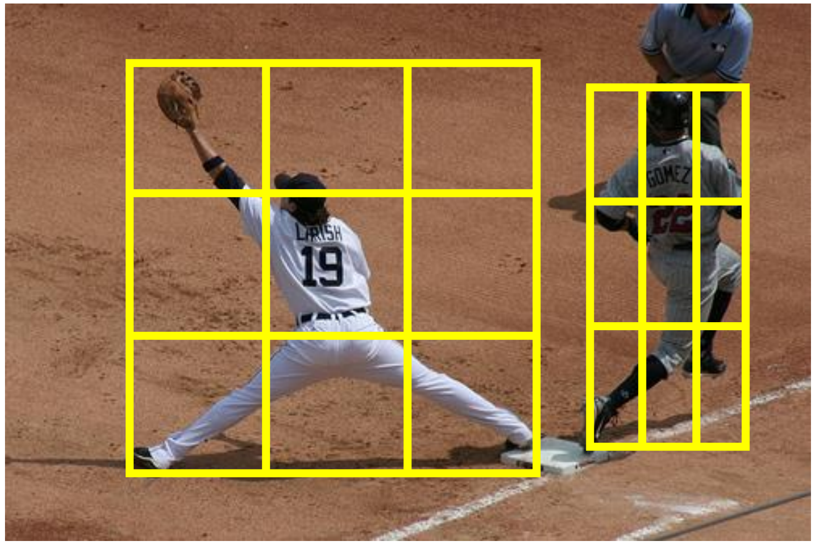
- Regular CNNs are inherently limited to model large unknown transformations
  - The limitation originates from the fixed geometric structures of CNN modules



regular convolution



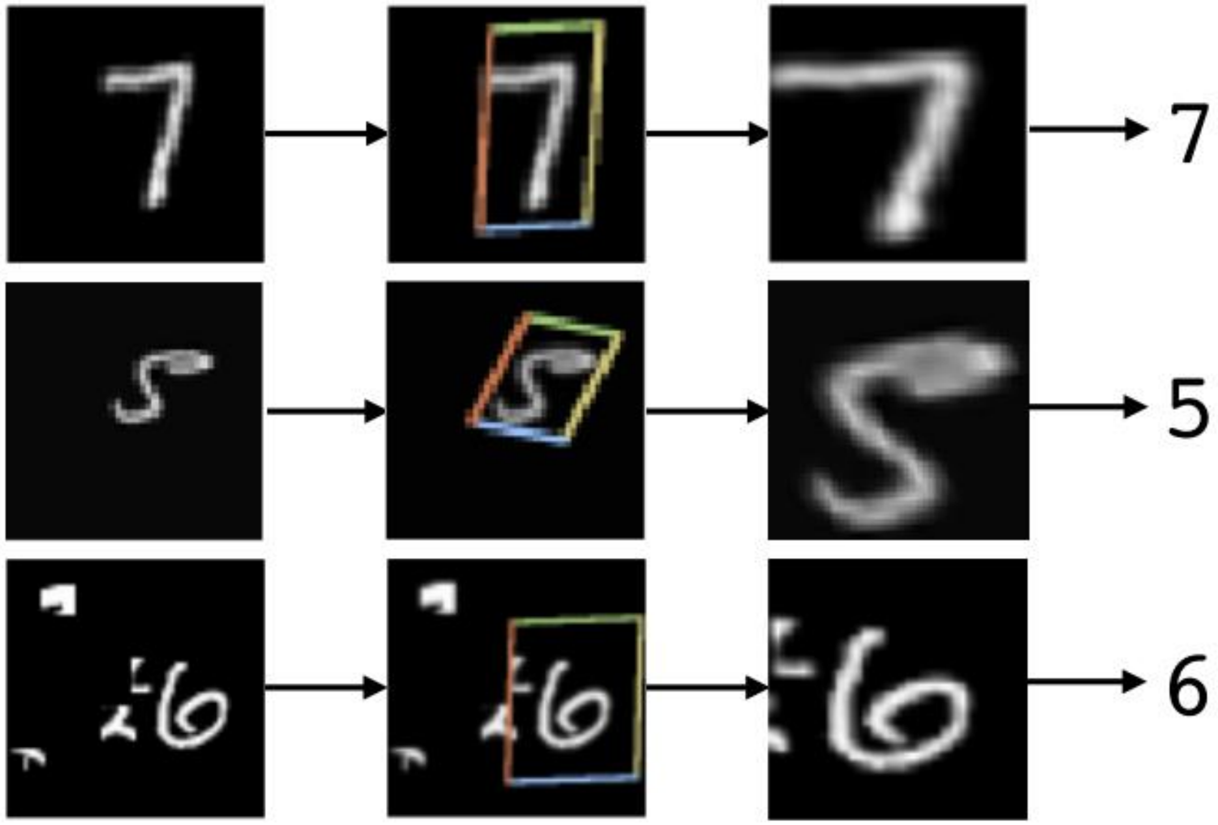
2 layers of regular convolution



regular RoI Pooling

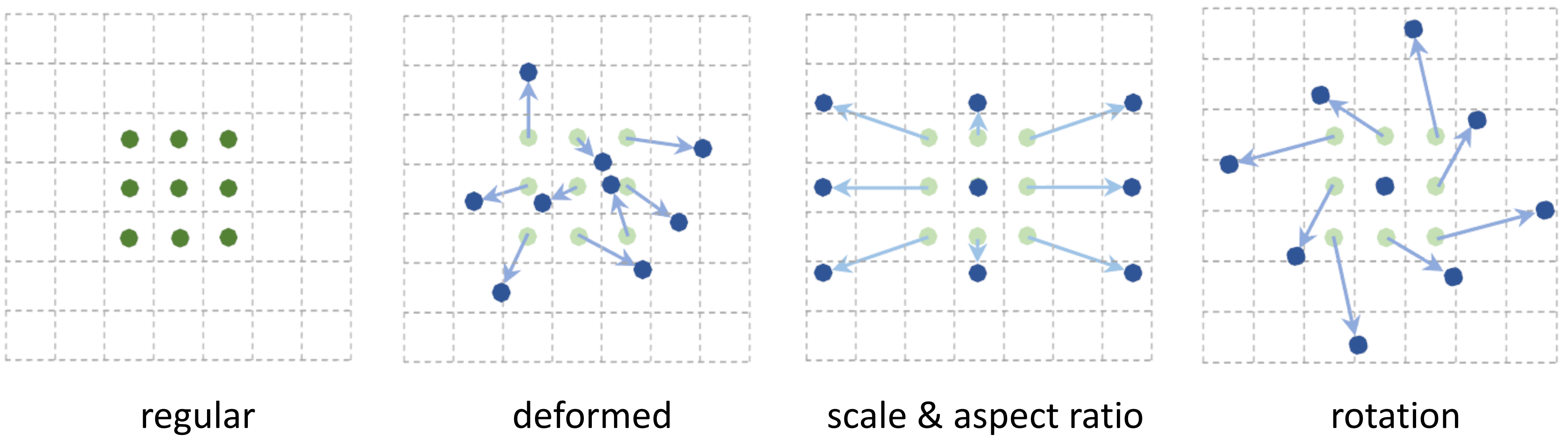
Spatial Transformer Networks

- Learning a global, parametric transformation on feature maps
  - Prefixed transformation family, infeasible for complex vision tasks

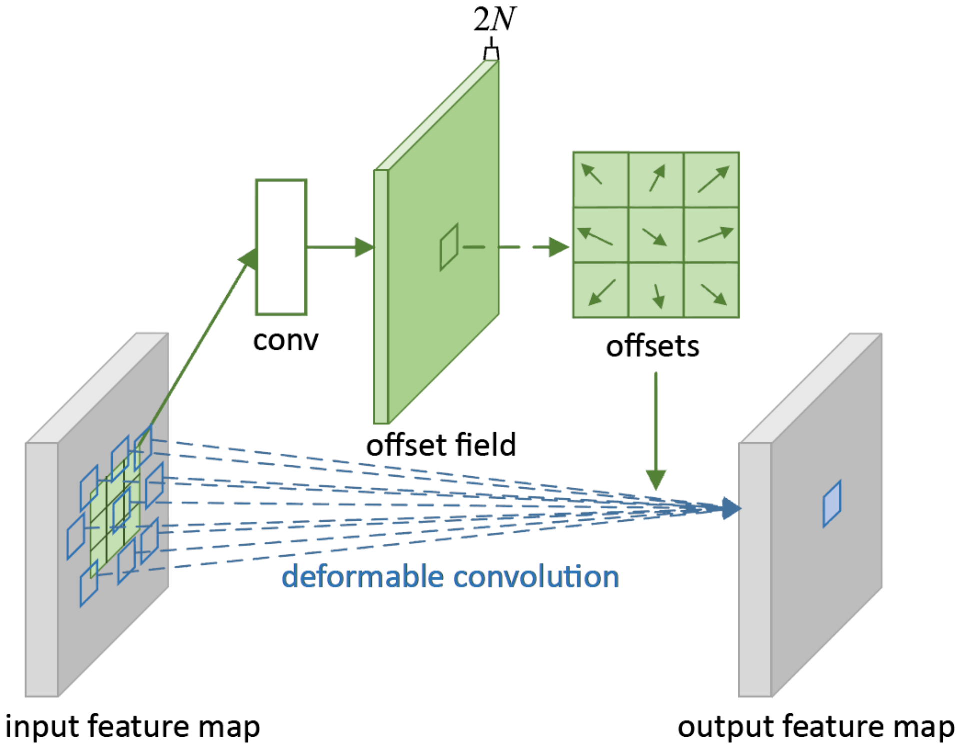


Deformable Convolution

- Local, dense, non-parametric transformation
  - Learning to deform the sampling locations in the convolution/RoI Pooling modules



Deformable Convolution



Regular convolution

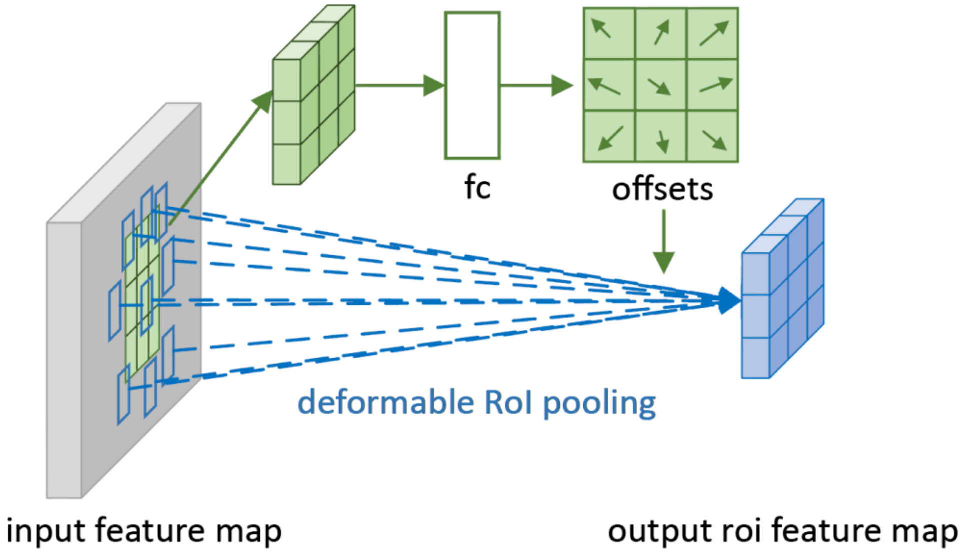
$$\mathbf{y}(\mathbf{p}_0) = \sum_{\mathbf{p}_n \in \mathcal{R}} \mathbf{w}(\mathbf{p}_n) \cdot \mathbf{x}(\mathbf{p}_0 + \mathbf{p}_n)$$

Deformable convolution

$$\mathbf{y}(\mathbf{p}_0) = \sum_{\mathbf{p}_n \in \mathcal{R}} \mathbf{w}(\mathbf{p}_n) \cdot \mathbf{x}(\mathbf{p}_0 + \mathbf{p}_n + \Delta\mathbf{p}_n)$$

where $\Delta\mathbf{p}_n$ is generated by a sibling branch of regular convolution

Deformable RoI Pooling



deformable RoI Pooling

Regular RoI pooling

$$\mathbf{y}(i,j) = \sum_{\mathbf{p} \in bin(i,j)} \mathbf{x}(\mathbf{p}_0 + \mathbf{p})/n_{ij}$$

Deformable RoI pooling

$$\mathbf{y}(i,j) = \sum_{\mathbf{p} \in bin(i,j)} \mathbf{x}(\mathbf{p}_0 + \mathbf{p} + \Delta\mathbf{p}_{ij})/n_{ij}$$

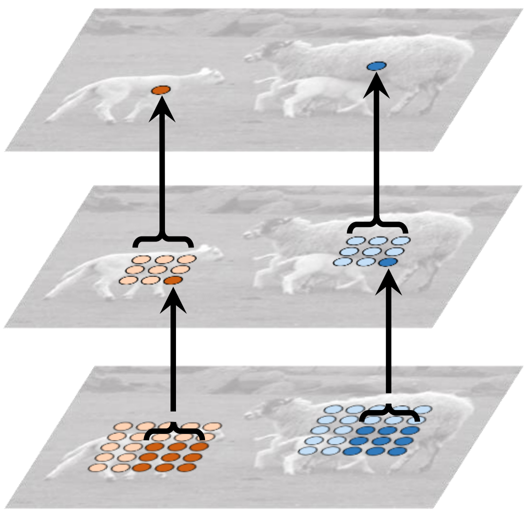
where $\Delta\mathbf{p}_{ij}$ is generated by a sibling fc branch

Deformable ConvNets

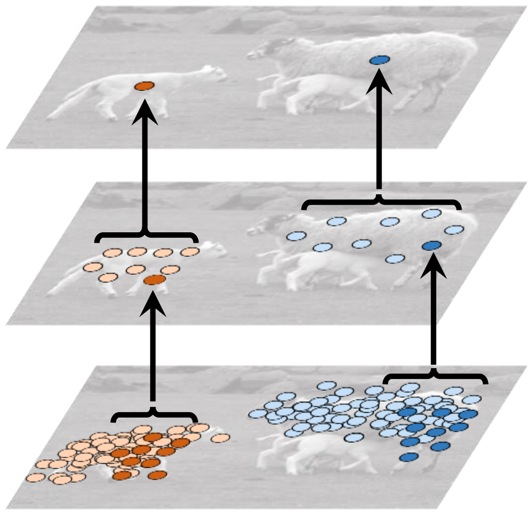
- Same input & output as the plain versions
  - Regular convolution -> deformable convolution
  - Regular RoI pooling -> deformable RoI pooling

- End-to-end trainable without additional supervision

Sampling Locations of Deformable Convolution

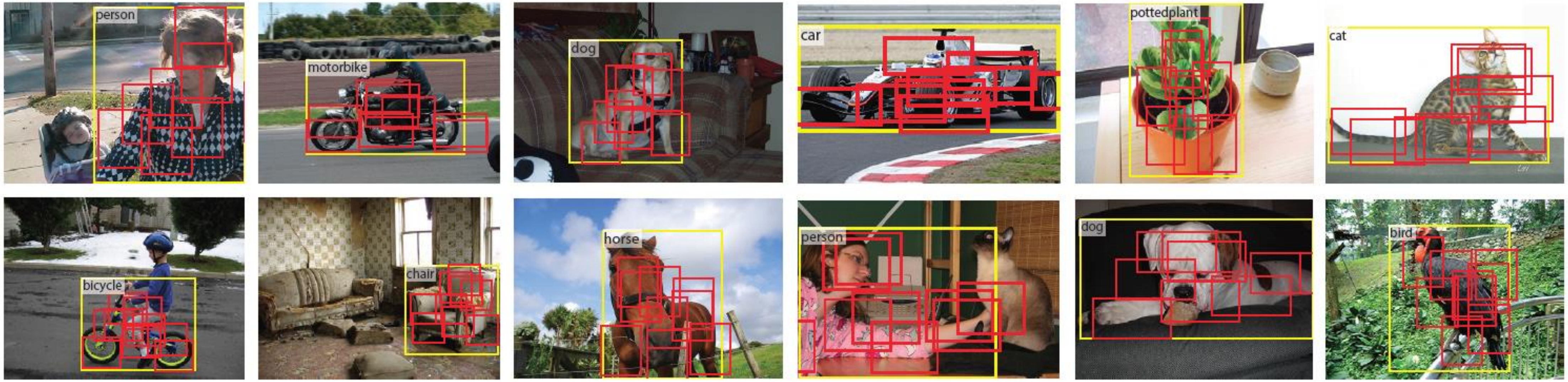


(a) standard convolution



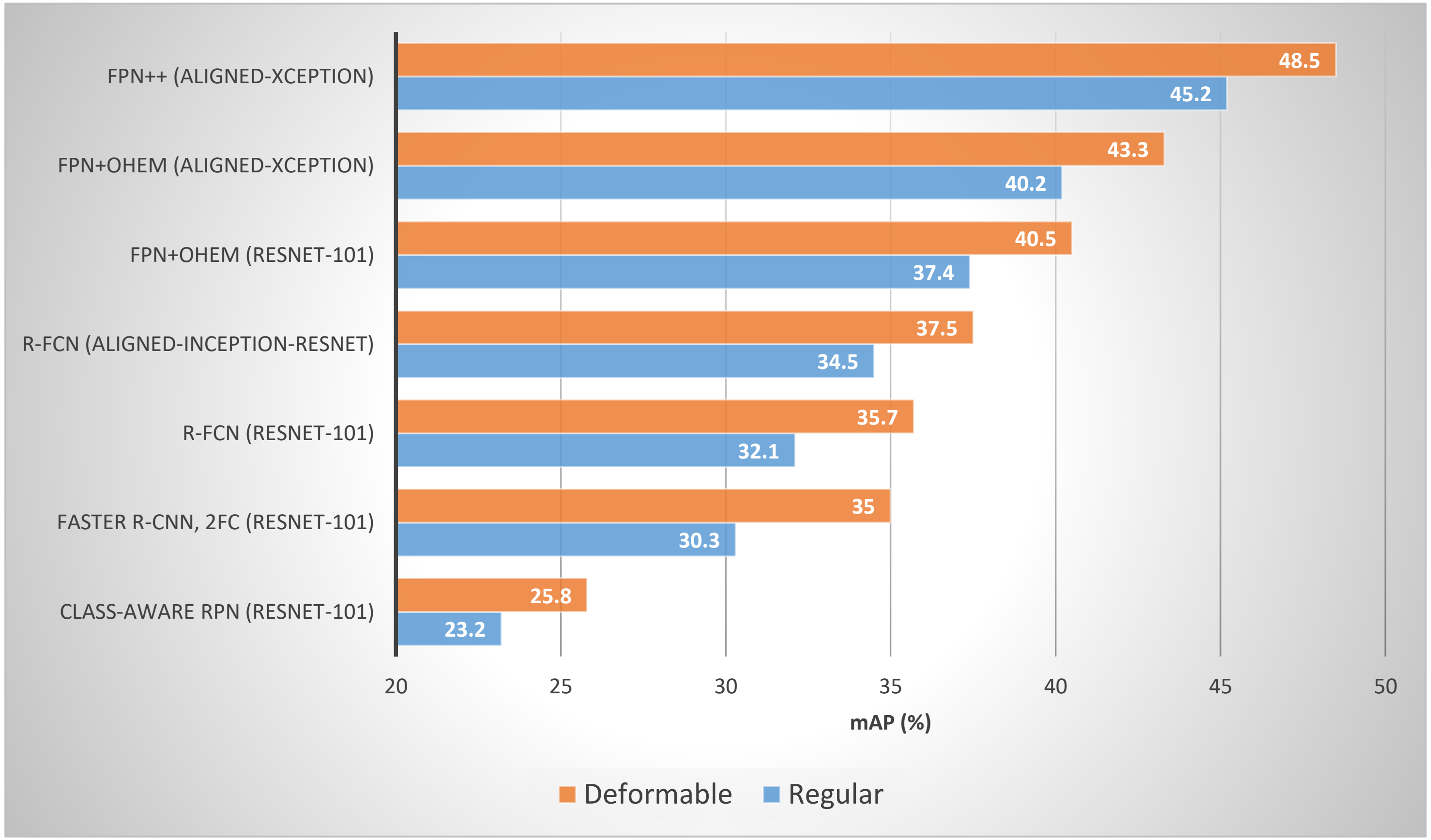
(b) deformable convolution

Part Offsets in Deformable RoI Pooling



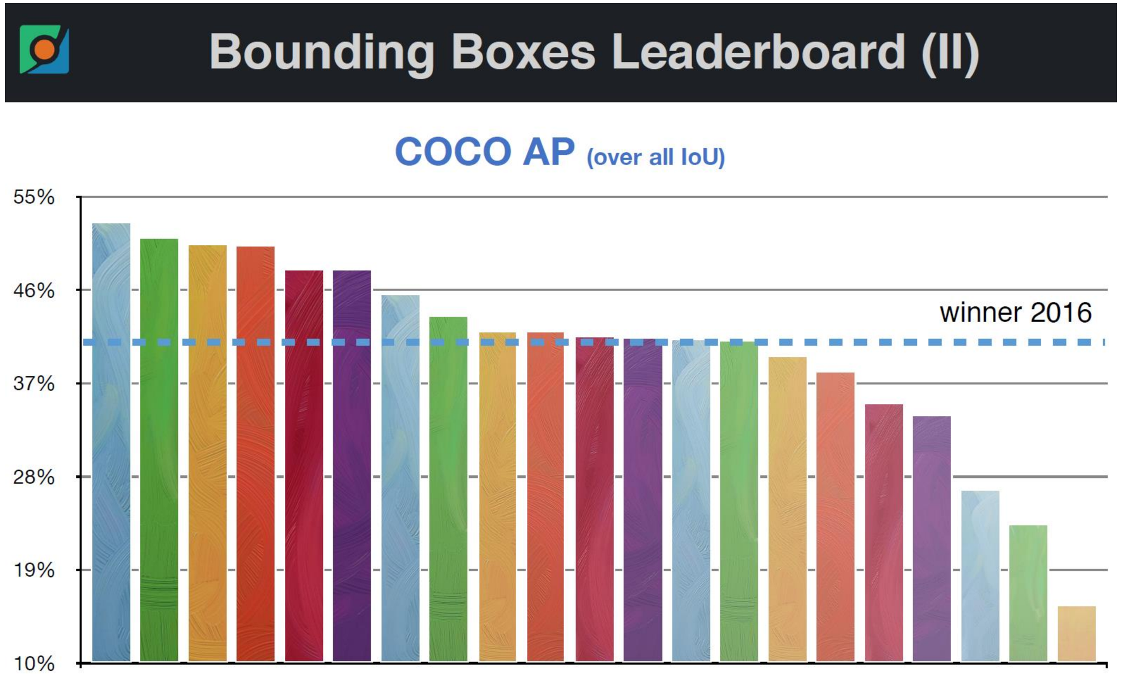
Object Detection on COCO (Test-dev)

- Deformable ConvNets v.s. regular ConvNets
  - Noticeable improvements for varies baselines
  - Marginal parameter & computation overhead



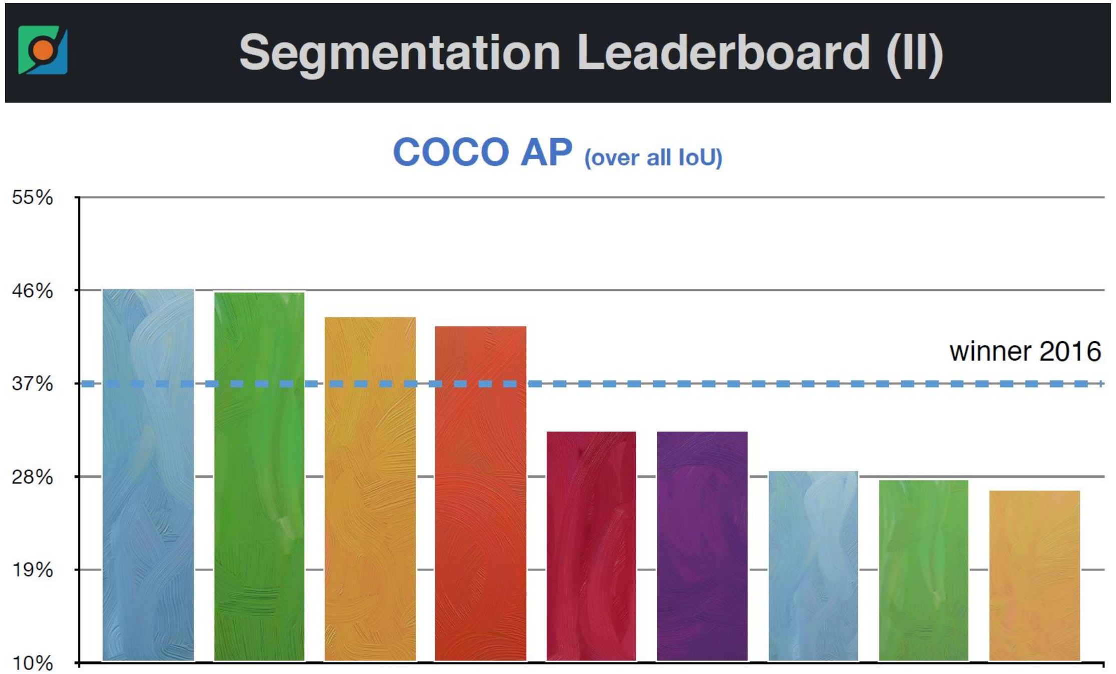
COCO Detection & Segmentation Challenge 2017

- Focus shifted from ImageNet to COCO in 2017
- Top-4 teams are quite close, surpassing others clearly



21 teams joined the competition
12 teams achieved better performance than last year's winner
4 teams > 50 AP

(*)



9 teams joined the competition
4 teams achieved better performance than last year's winner
4 teams > 40 AP

(*)

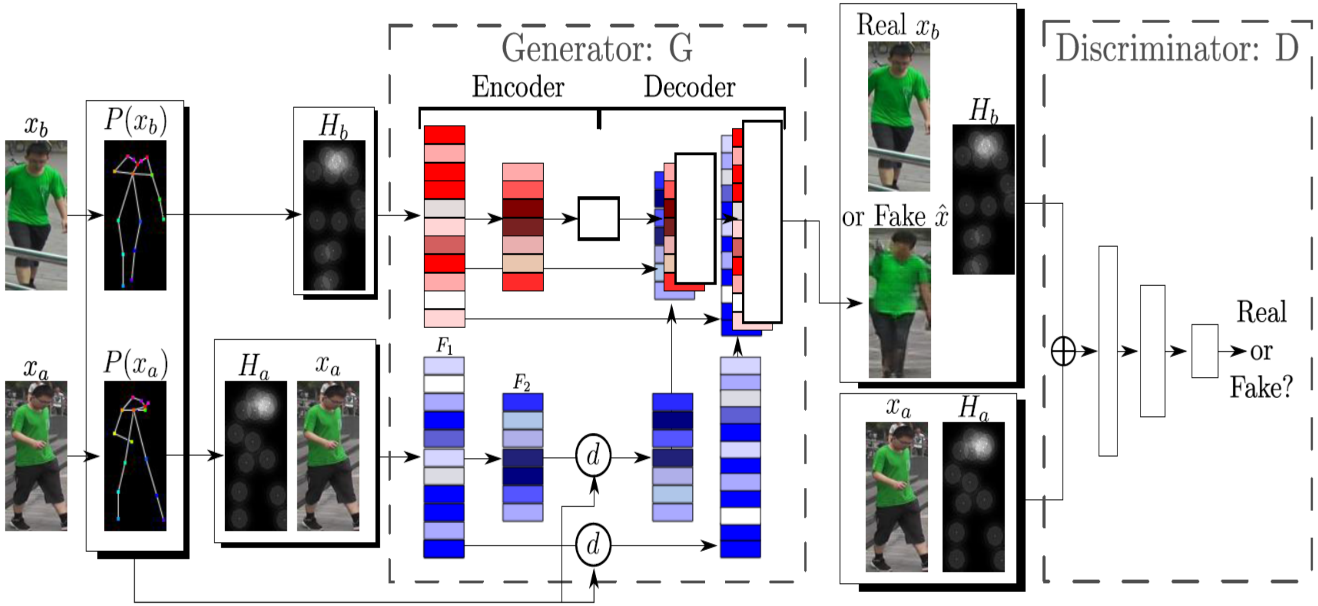
COCO Detection & Segmentation Challenge 2017

- Few tricks and hacks are adopted by MSRA and FAIR team
- Our accuracy is on par with FAIR team, at much smaller model size
- Deformable ConvNets are also adopted by other teams

Team	BBox	Segmentation	Tricks & Hacks	Model Ensembled	Utilize of Deformable CNNs
Megvii (Face++)	1st	2nd	Many	Unknown	Unknown
Ucenter (SenseTime)	2nd	1st	Many	Unknown	***Yes***
MSRA	3rd	4th	***Few***	**6**	***Yes***
FAIR	4th	3rd	***Few***	**30**	No

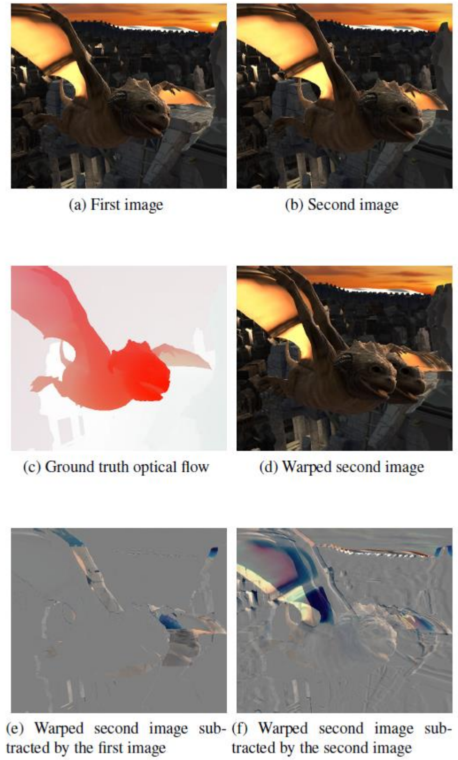
Deformable ConvNets Extensions I

- Deformable GANs



[Siarohin et al. Arxiv Tech Report, 2017.]

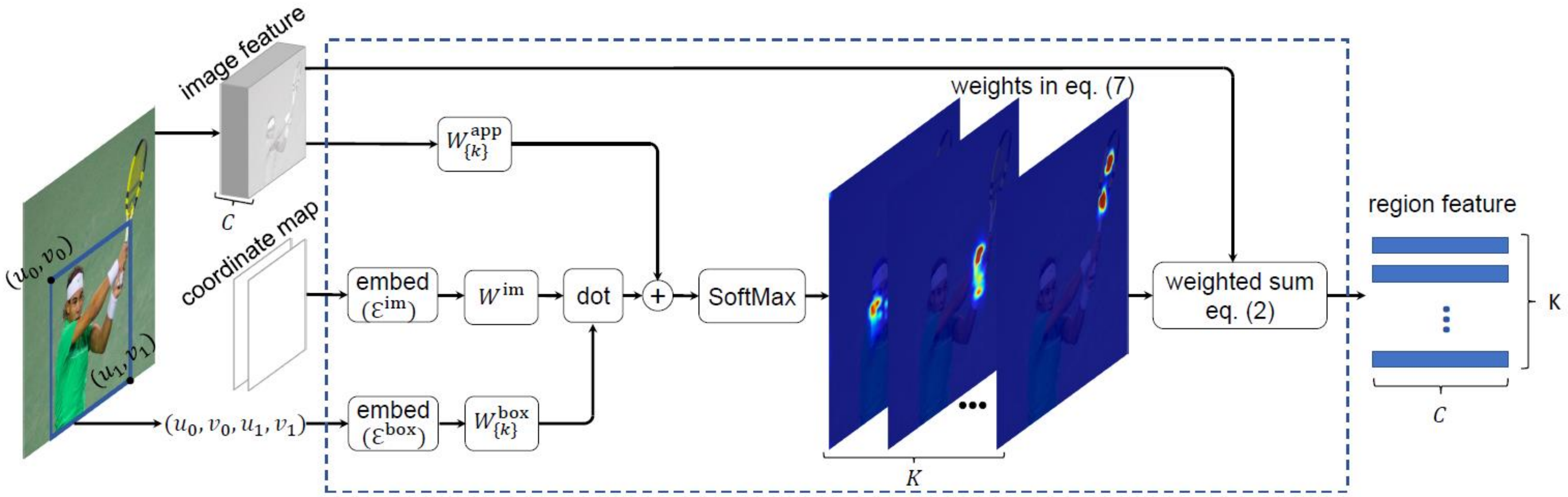
- Deformable volume network for flow estimation

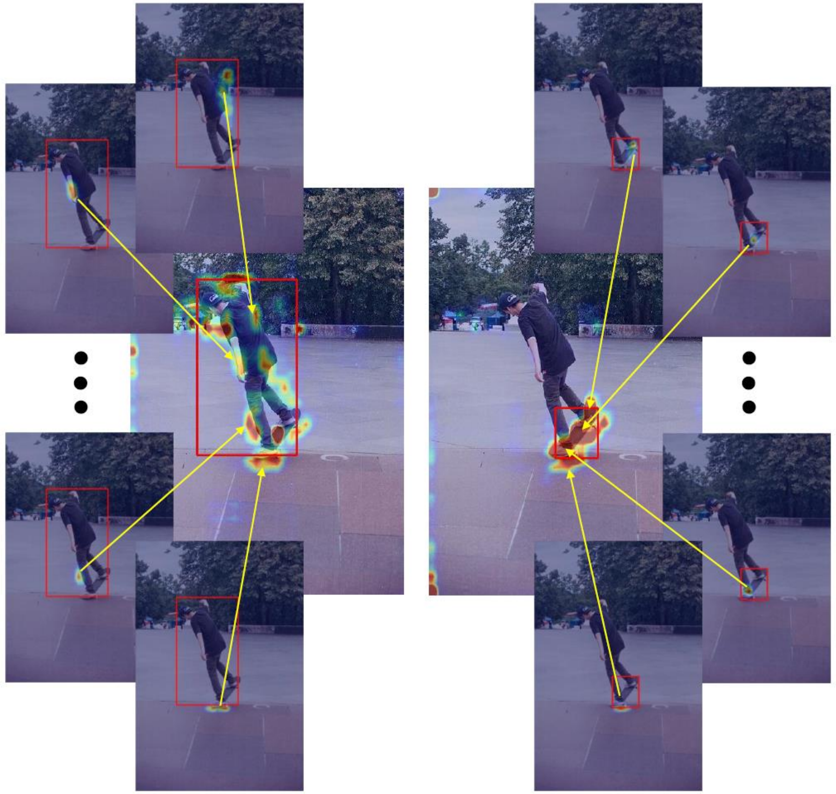


[Lu et al. Arxiv Tech Report, 2018.]

Deformable ConvNets Extensions II

- Fully learnable region feature extraction
  - Deformed regular grid, offset learning -> Free-form shape, attention weight learning





[Gu et al. Arxiv Tech Report, 2018.]

Outline

- R-FCN and its extensions
- Deformable ConvNets and its extensions
- Video object detection
- Summary

Per-frame recognition in video is problematic

High Computational Cost

Infeasible for practical needs

Task	Image Size	ResNet-50	ResNet-101
Detection	1000x600	6.27 fps	4.05 fps
Segmentation	2048x1024	2.24 fps	1.52 fps

FPS: frames per second
(NVIDIA K40 and Intel Core i7-4790)

Deteriorated Frame Appearance

Poor feature and recognition accuracy

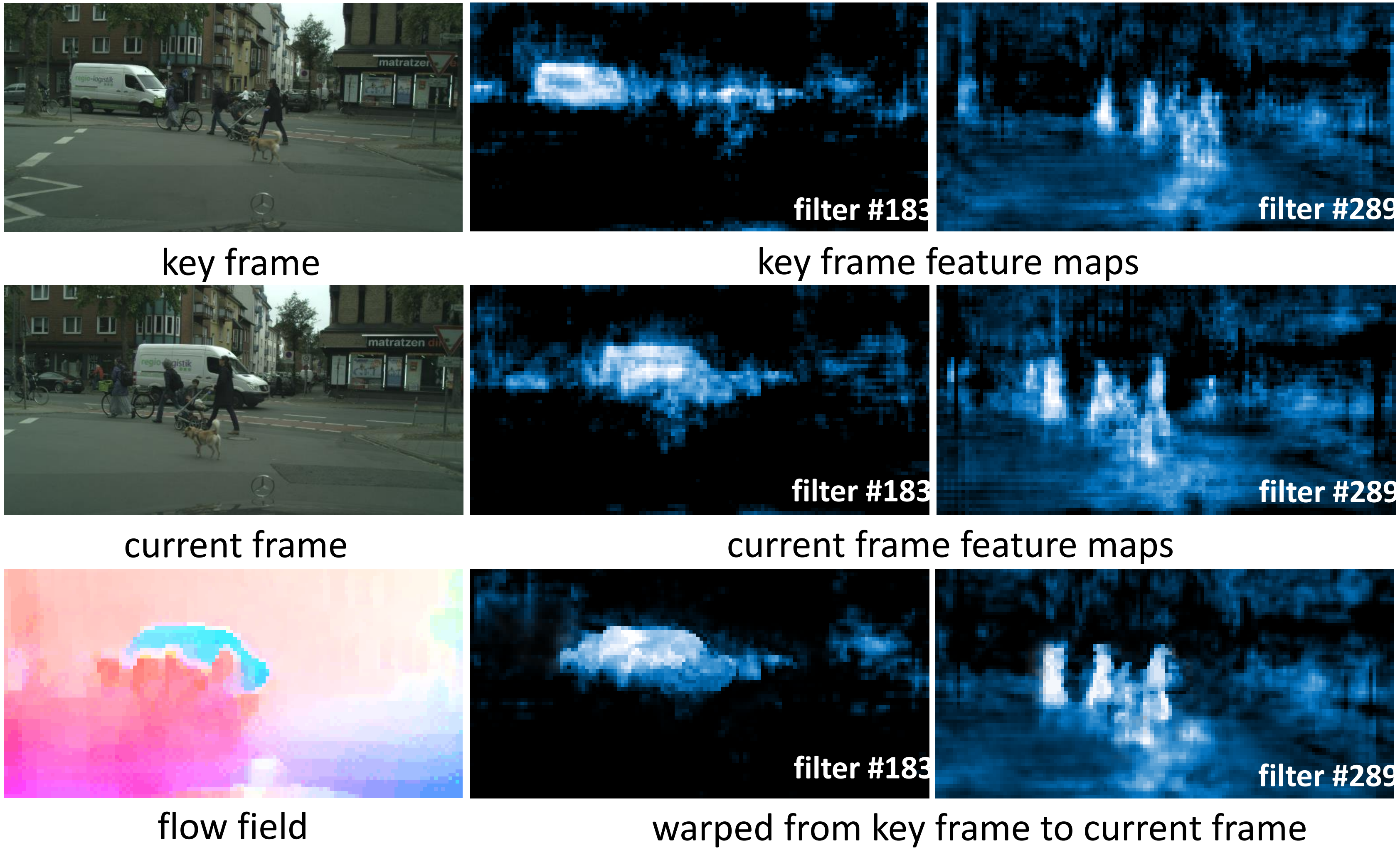
motion blur

part occlusion

rare poses

Key idea

- Flow-guided feature propagation & aggregation



Powering the winner of ImageNet VID 2017

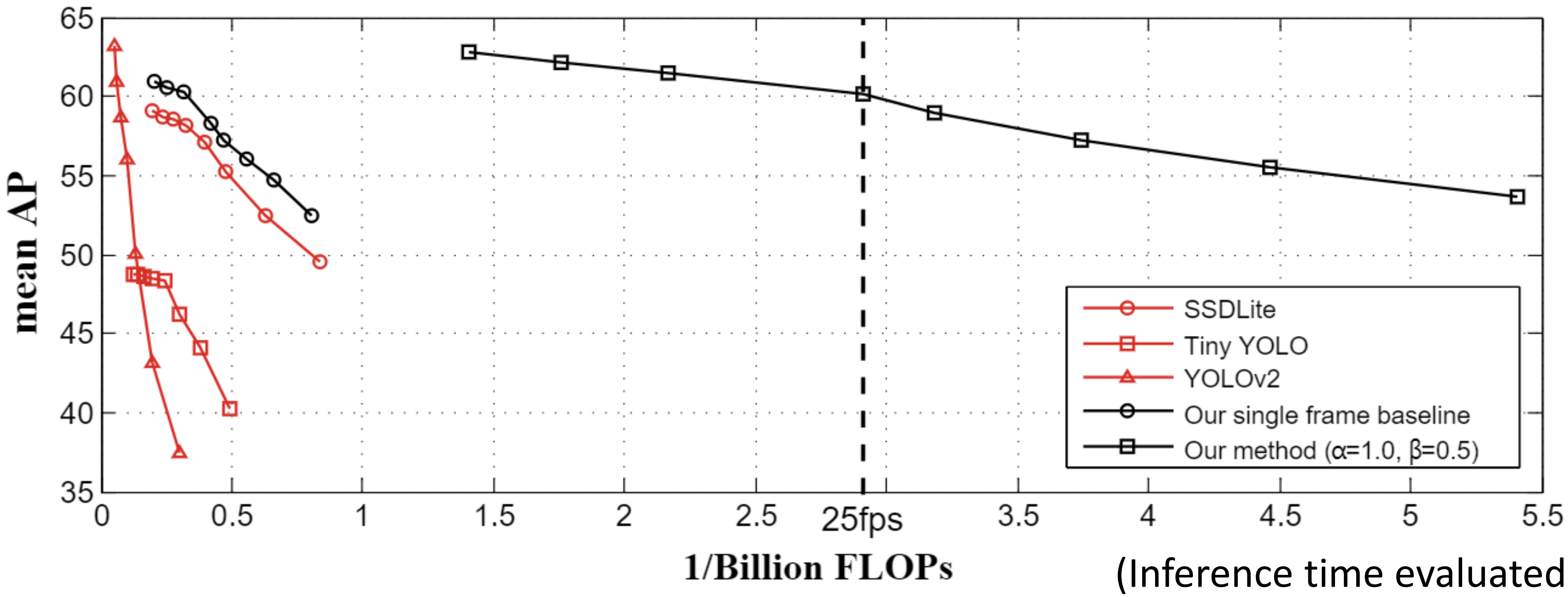
Team name	Entry description	**Number of object categories won**	mean AP
IC&USYD	provide_submission3	15	0.817265
IC&USYD	provide_submission1	6	0.808847
IC&USYD	provide_submission2	4	0.818309
NUS-Qihoo-UIUC_DPNs (VID)	no_extra + seq + mca + mcs	3	0.757772
NUS-Qihoo-UIUC_DPNs (VID)	no_extra + seq + vcm + mcs	1	0.757853
NUS-Qihoo-UIUC_DPNs (VID)	Faster RCNN + Video Context	1	0.748493
THU-CAS	merge-new	0	0.730498
THU-CAS	old-new	0	0.728707
THU-CAS	new-new	0	0.691423
GoerVision	Deformable R-FCN single model+ResNet101	0	0.669631
GoerVision	Ensemble 2 model, use ResNet101 as foundamental classification network and deformable R-FCN to detect video frames, multi-scale testing	0	0.665693
GoerVision	o train the video objectWe use the ResNet101 and Deformable R-FCN for the detection.	0	0.655686
GoerVision	Using R-FCN to detect video object, multi scale testing applied.	0	0.646965
FACEALL_BUPT	SSD based on Resnet101 networks	0	0.195754

[top]

IC&USYD	Jiankang Deng(1), Yuxiang Zhou(1), Baosheng Yu(2), Zhe Chen(2), Stefanos Zafeiriou(1), Dacheng Tao(2), (1)Imperial College London, (2)University of Sydney	Flow acceleration[1,2] is used. Final scores are adaptively chosen between the detector and tracker. [1] Deep Feature Flow for Video Recognition Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. [2] Flow-Guided Feature Aggregation for Video Object Detection, Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Arxiv tech report, 2017.

Towards High Performance Video Object Detection for Mobiles

- Accurate, real-time video object detection on mobiles for the first time
- An order faster than previous fastest object detectors with on par accuracy



(Inference time evaluated on Huawei Mate 8)

Outline

- R-FCN and its extensions
- Deformable ConvNets and its extensions
- Video object detection
- Summary

Summary

- General object detection is still an open, unsolved, fundamental vision problem
  - Recognition of objects with large appearance variations
  - Low recognition latency on mobile devices
  - Panoramic scene understanding

- Careful investigation and prototyping is necessary in application in products