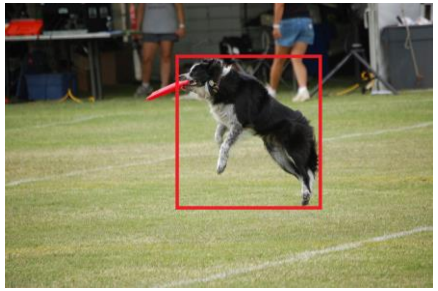
High-Resolution Networks

Jingdong Wang
Senior Principal Research Manager
Microsoft Research, Beijing, China

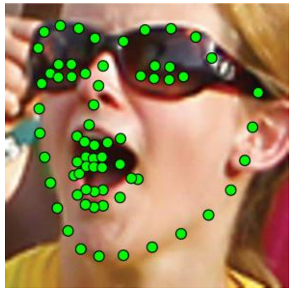
Convolutional neural networks are good at representation learning

Image classification



Object detection

Semantic segmentation



Face alignment

Pose estimation

……

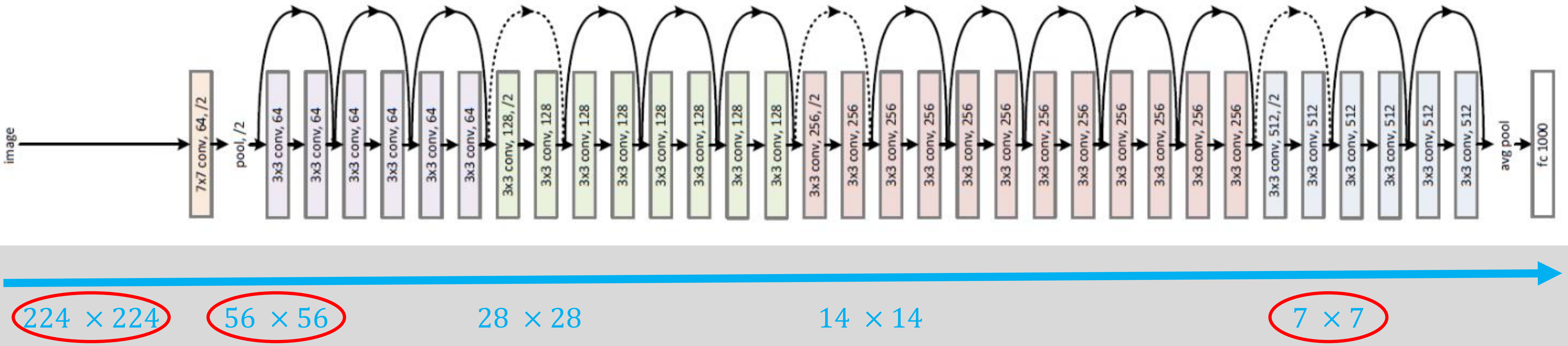

Low-resolution representation learning
High-resolution representation learning
image classification
region-level recog.
pixel-level recog.
global
position-sensitive


Low-resolution representation learning
High-resolution
image classification
global

Low-resolution representation learning

Classification networks: connect the convolutions in *series* from high resolution to low resolution

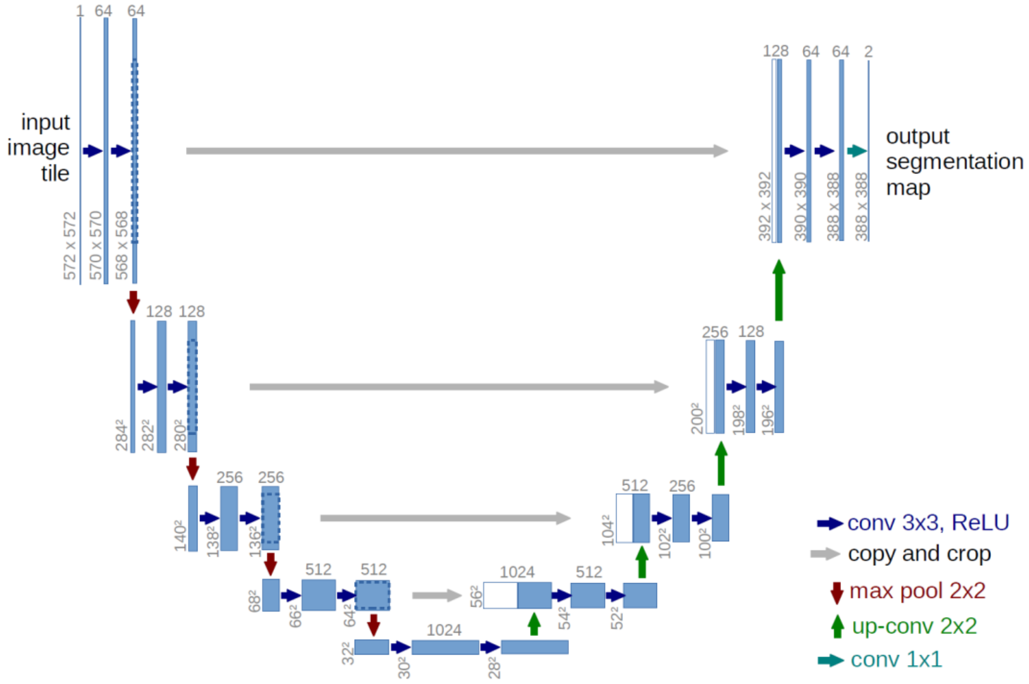
Standard design and followed by AlexNet, VGGNet, GoogleNet, ResNet, DensetNet



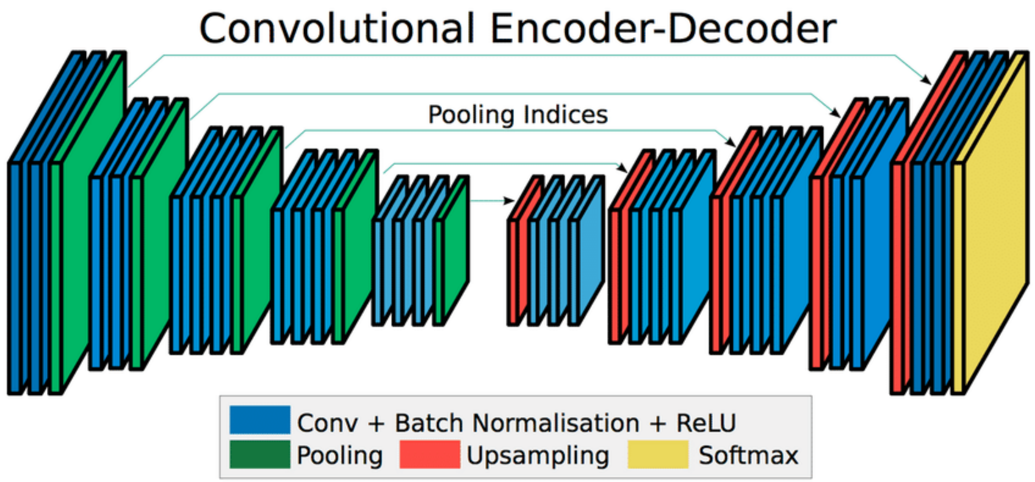
ResNet


High-resolution representation learning
region-level recog.
pixel-level recog.
position-sensitive

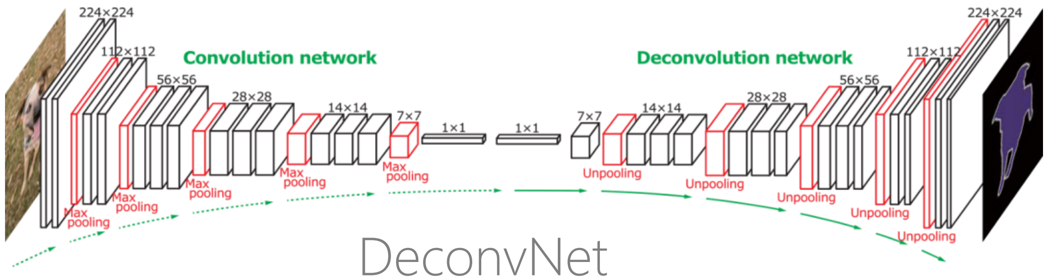
Previous high-resolution representation learning



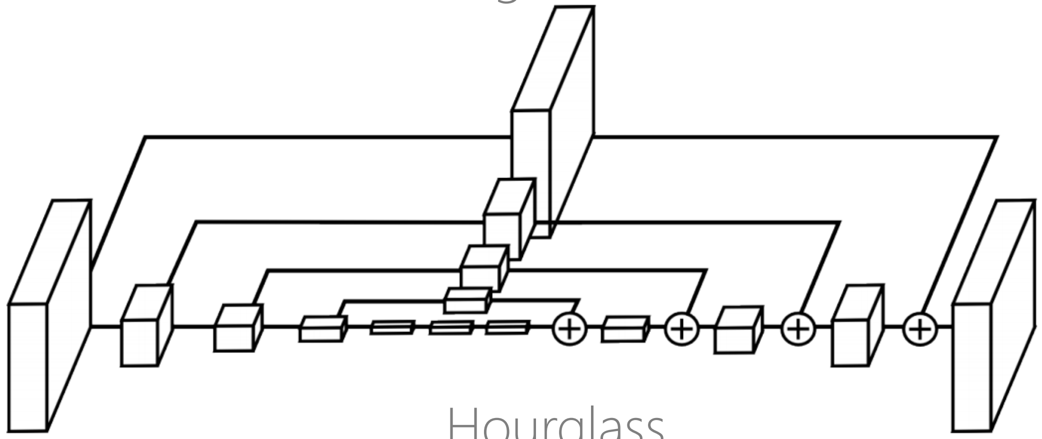
U-Net



SegNet



DeconvNet



Hourglass

Previous SOTA solutions: look different, essentially the same

Previous high-resolution representation learning

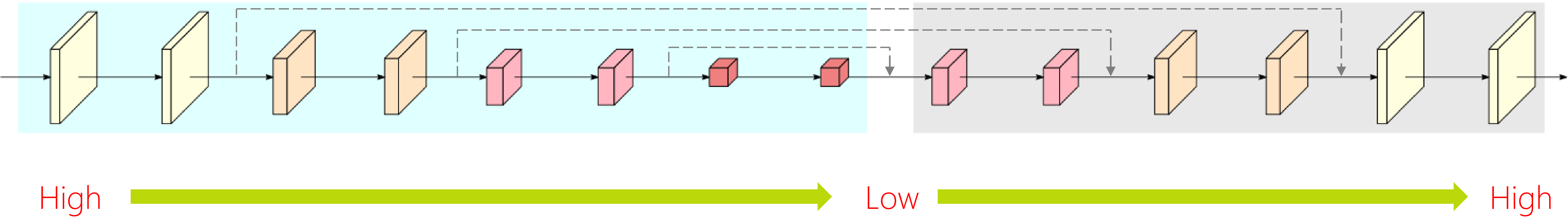
Essentially, the previous methods **remediate/extend** classification networks (e.g., ResNet)

- ❑ **Stage** 1: compute low-resolution representations using a classification network
- ❑ **Stage** 2: recover high resolution from low resolution by sequentially-connected convolutions

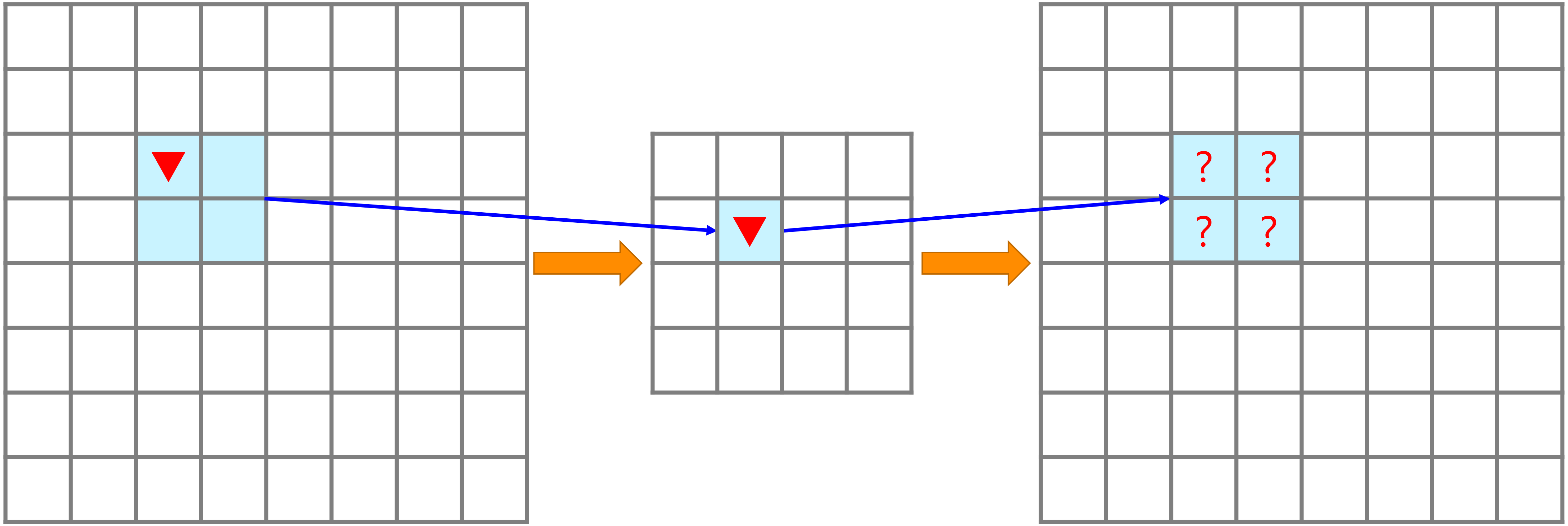
Previous high-resolution representation learning

Essentially, the previous methods **remediate/extend** classification networks (e.g., ResNet)

- ❑ **Stage** 1: compute low-resolution representations using a classification network
- ❑ **Stage** 2: recover high resolution from low resolution by sequentially-connected convolutions



High→low→high leads to position-sensitivity loss



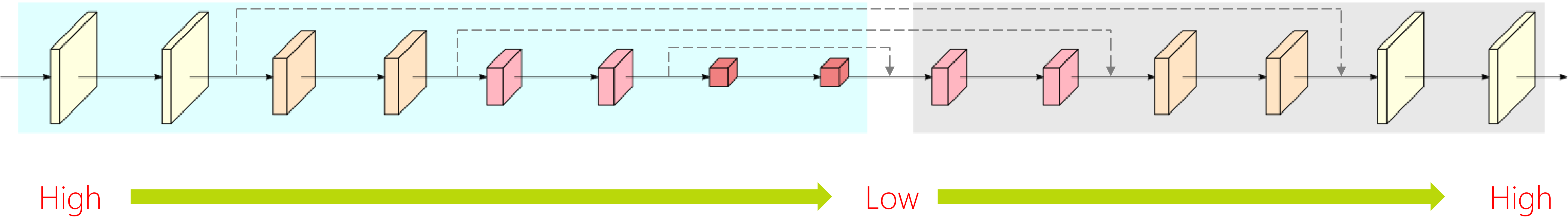
8×8 (high) 4×4 (low) 8×8 (high)

Previous high-resolution representation learning

Essentially, the previous methods **remediate/extend** classification networks (e.g., ResNet)

- ❑ **Stage** 1: compute low-resolution representations using a classification network
- ❑ **Stage** 2: recover high resolution from low resolution by sequentially-connected convolutions

☹ The position-sensitivity of the representation is weak



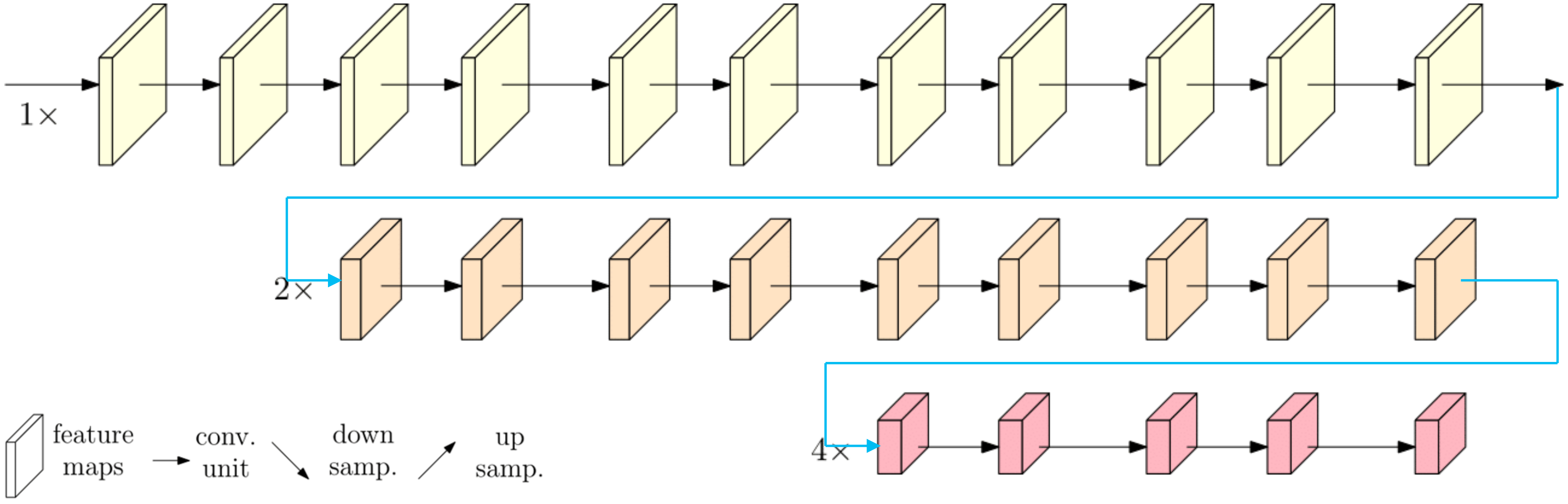
Our work: High-resolution network (HRNet)

- Learn high-resolution representations with stronger position sensitivity
- Design from scratch instead of from classification networks
- Maintain high resolution representations through the whole network other than recovering from low resolution

Ke Sun, Bin Xiao, Dong Liu, Jingdong Wang: Deep High-Resolution Representation Learning for Human Pose Estimation. CVPR 2019
Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao: Deep High-Resolution Representation Learning for Visual Recognition (submitted to TPAMI)

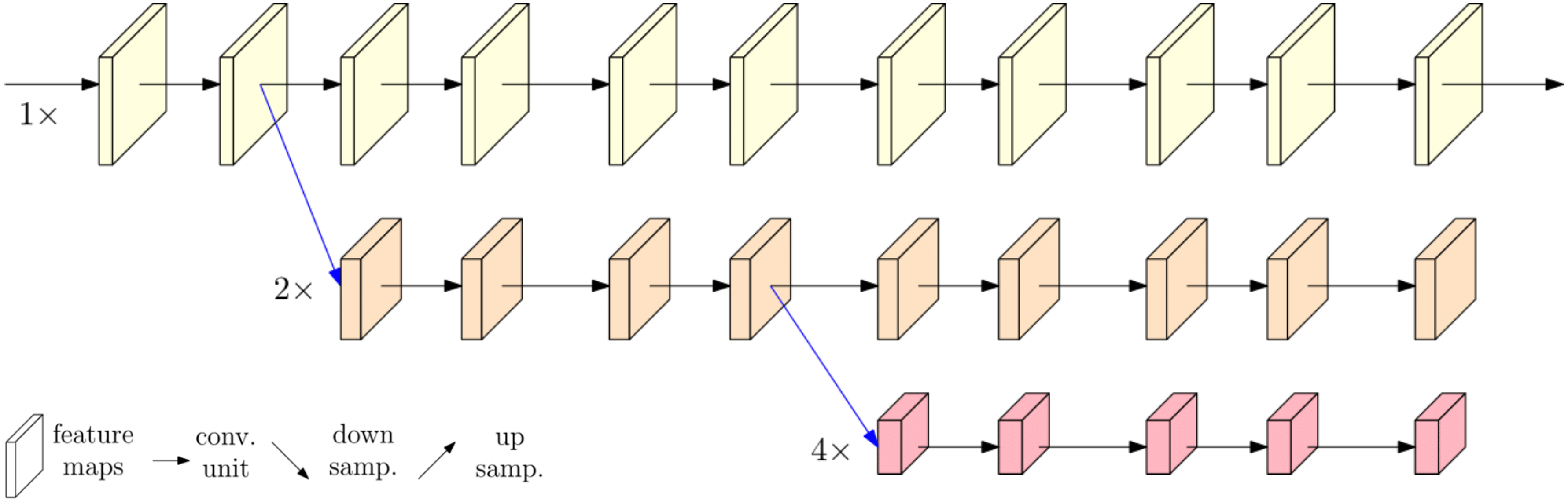
Previous low-resolution networks

Connect multi-resolution convolutions in *series* from high to low



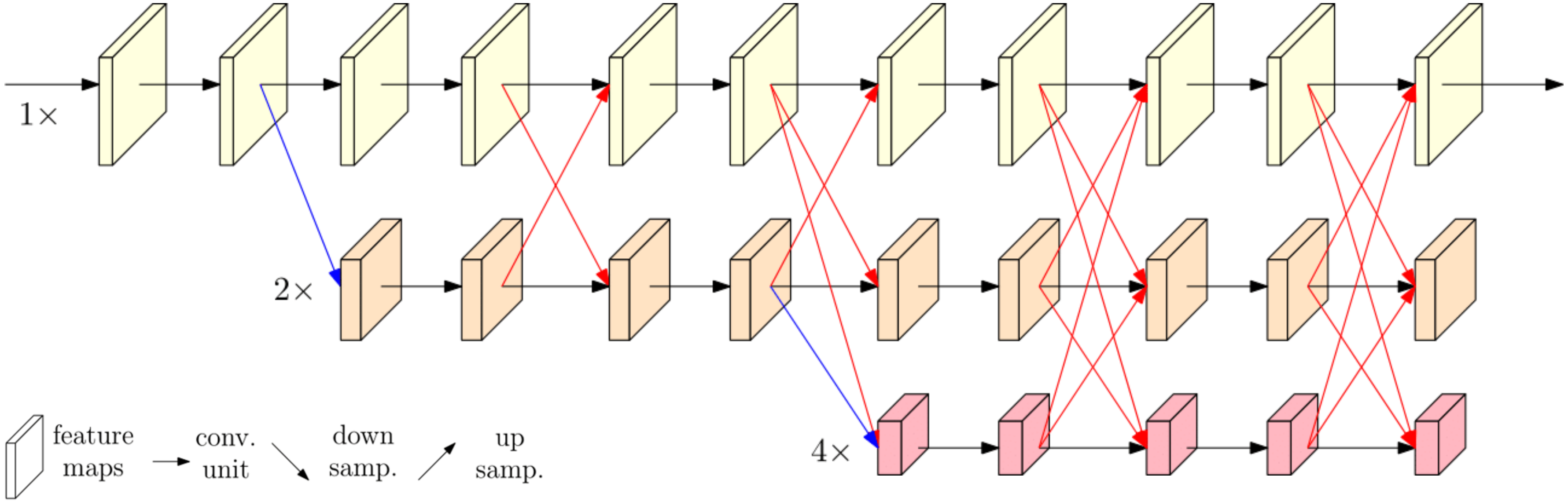
HRNet: high-resolution representation learning

High-resolution networks (HRNet): Connect multi-resolution convolutions in ***parallel***



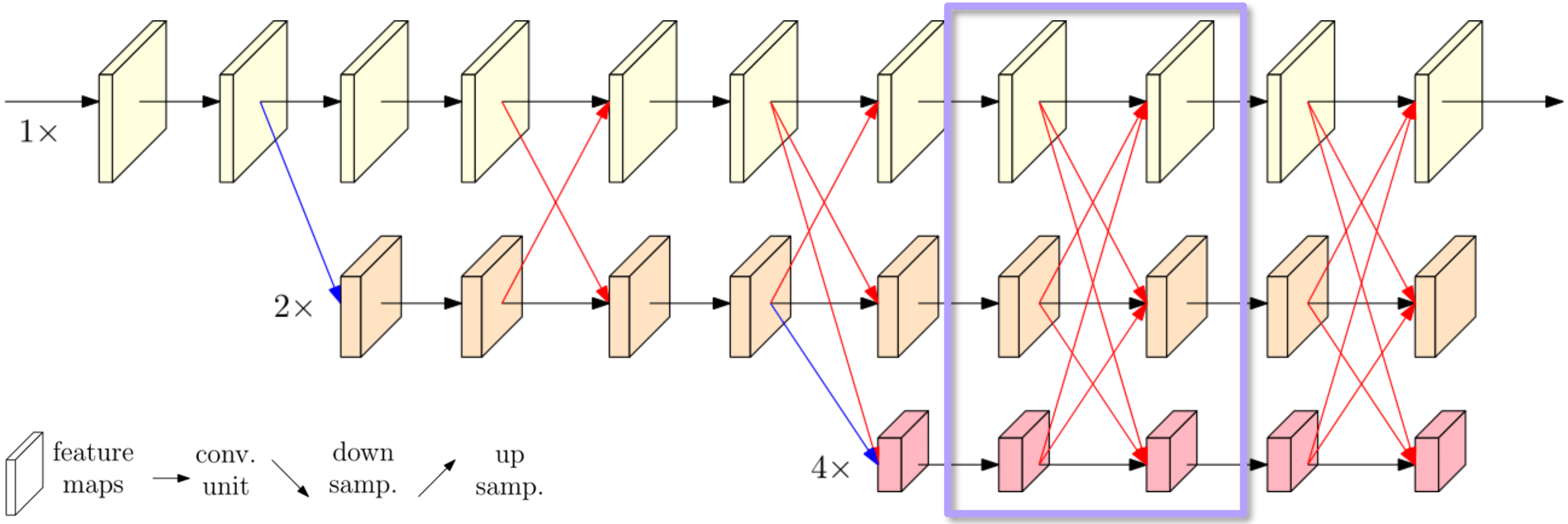
HRNet: high-resolution representation learning

High-resolution networks (HRNet): Connect multi-resolution convolutions in ***parallel*** with ***repeated fusions***

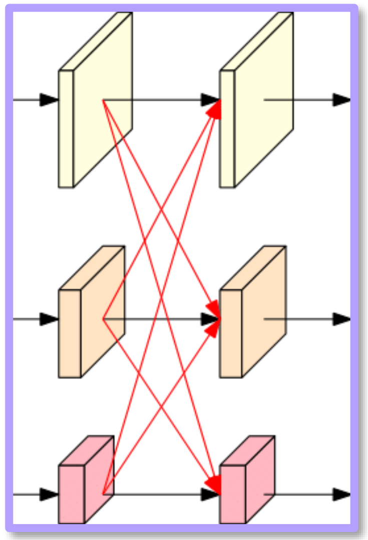


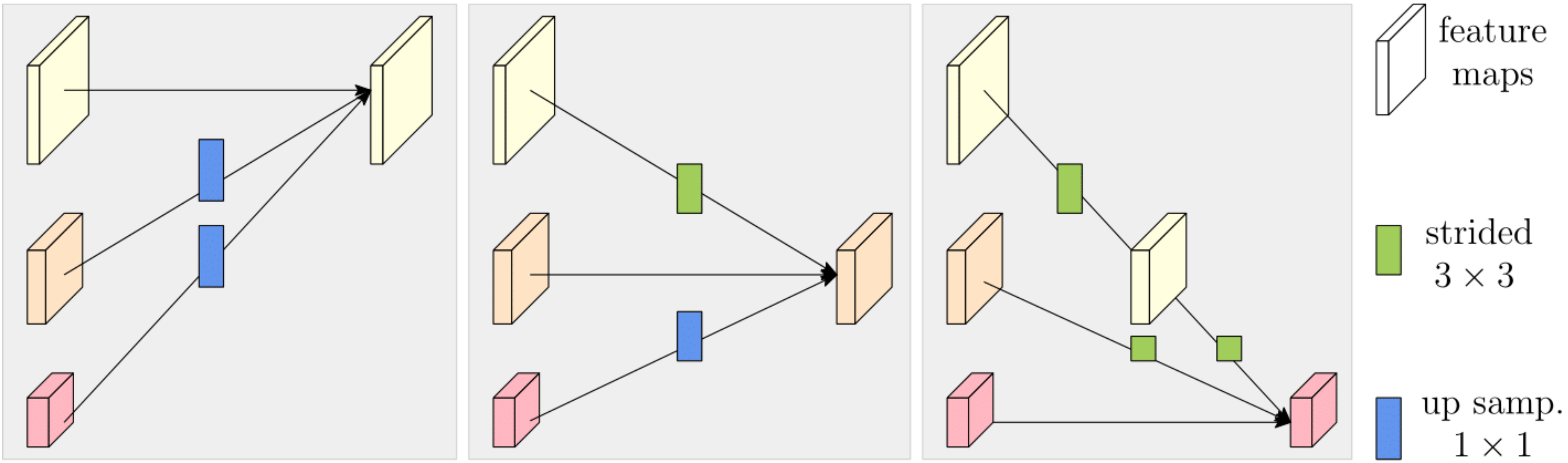
HRNet: high-resolution representation learning

High-resolution networks (HRNet): Connect multi-resolution convolutions in ***parallel*** with ***repeated fusions***



Across-resolution fusion





Down-sample: $\mathbf{stride} - 2\ 3 \times 3$
Up-sample: $\mathbf{bilinear} + 1 \times 1$

Fundamental architecture changes

parallel

- Connect high-to-low resolution convolutions in ~~*series*~~

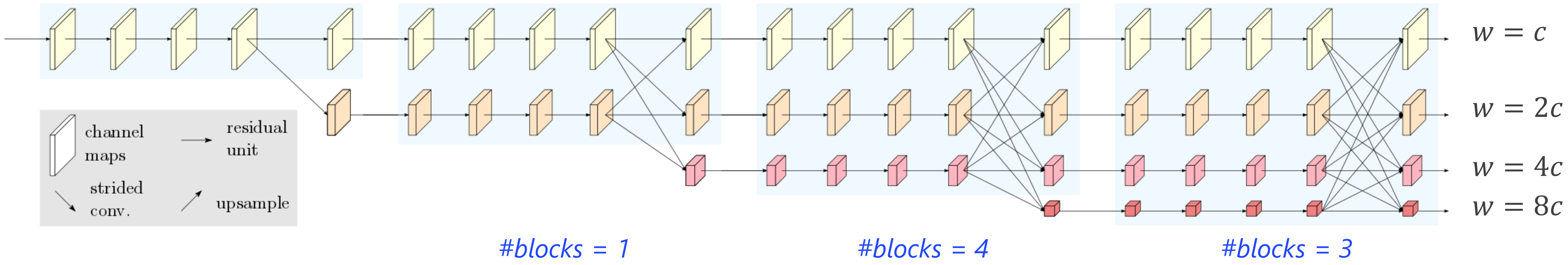
Maintain through the whole process

- ~~*Recover*~~ high-resolution representations ~~*from low-resolution representations*~~

- Repeat fusions across resolutions to strengthen high- & low-resolution representations

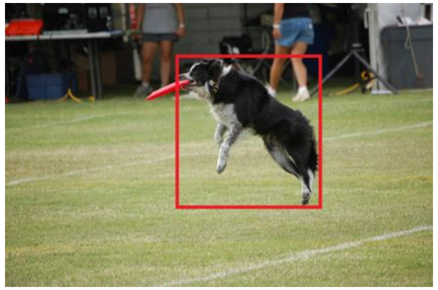
HRNet can learn *high-resolution* representations with *strong position sensitivity*

HRNet instantiation



- ❑ Fix the depth and change the width for tuning the capacity.
- ❑ The width (e.g., $c = 32, 48$) is much smaller than the ResNet (256).
- ❑ The parameter and computation complexities are similar to ResNet-based methods.

Human pose estimation



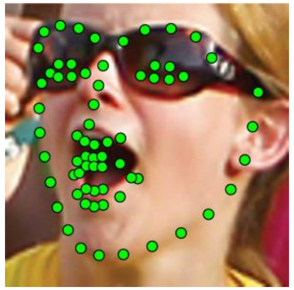


Image classification

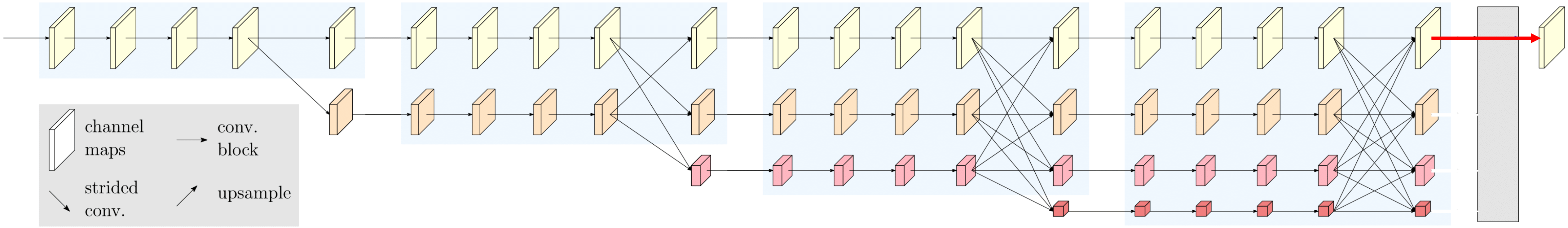
Object detection

Semantic segmentation

Face alignment

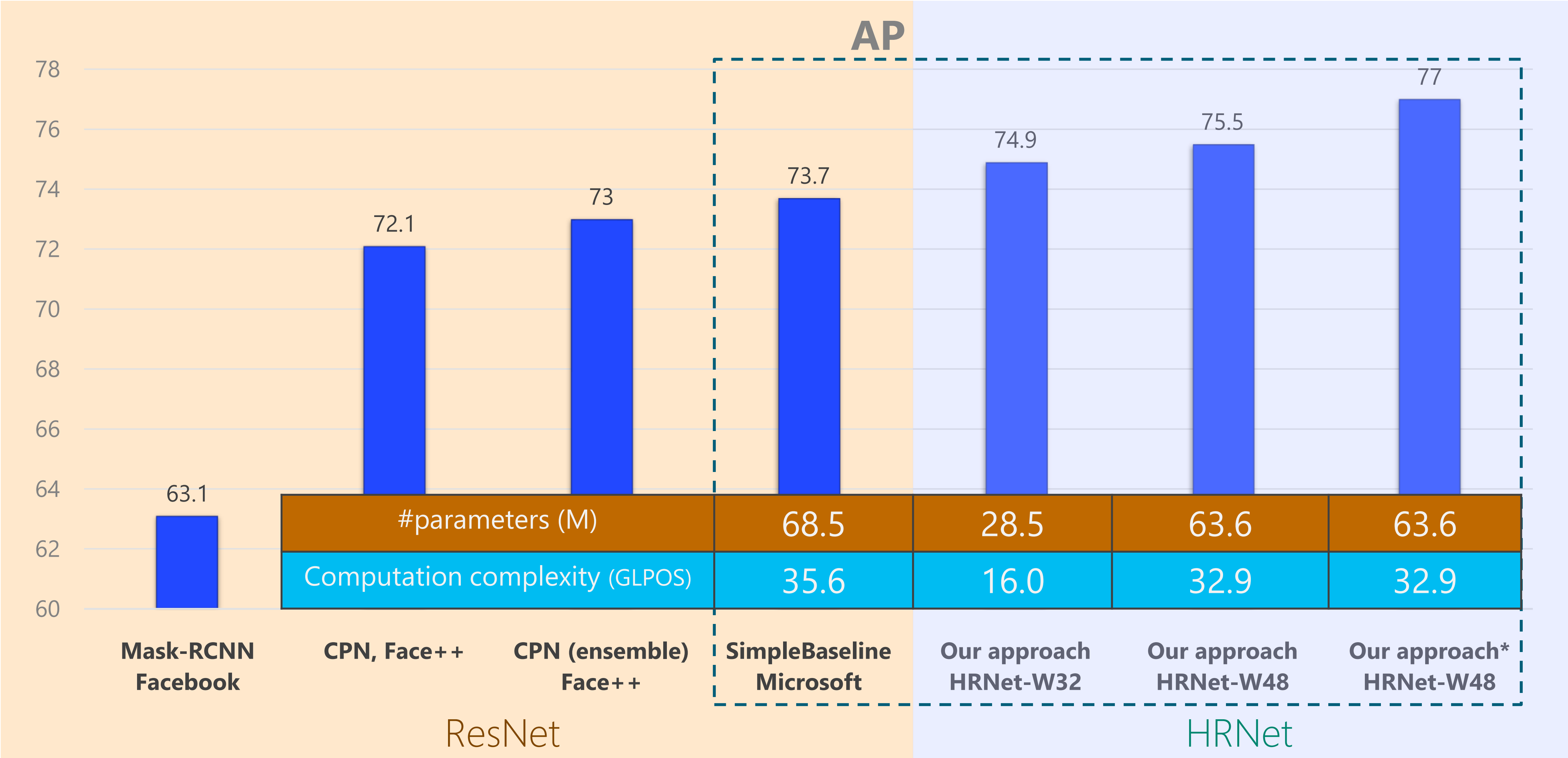
Pose estimation

HRNet for human pose estimation



COCO human pose estimation

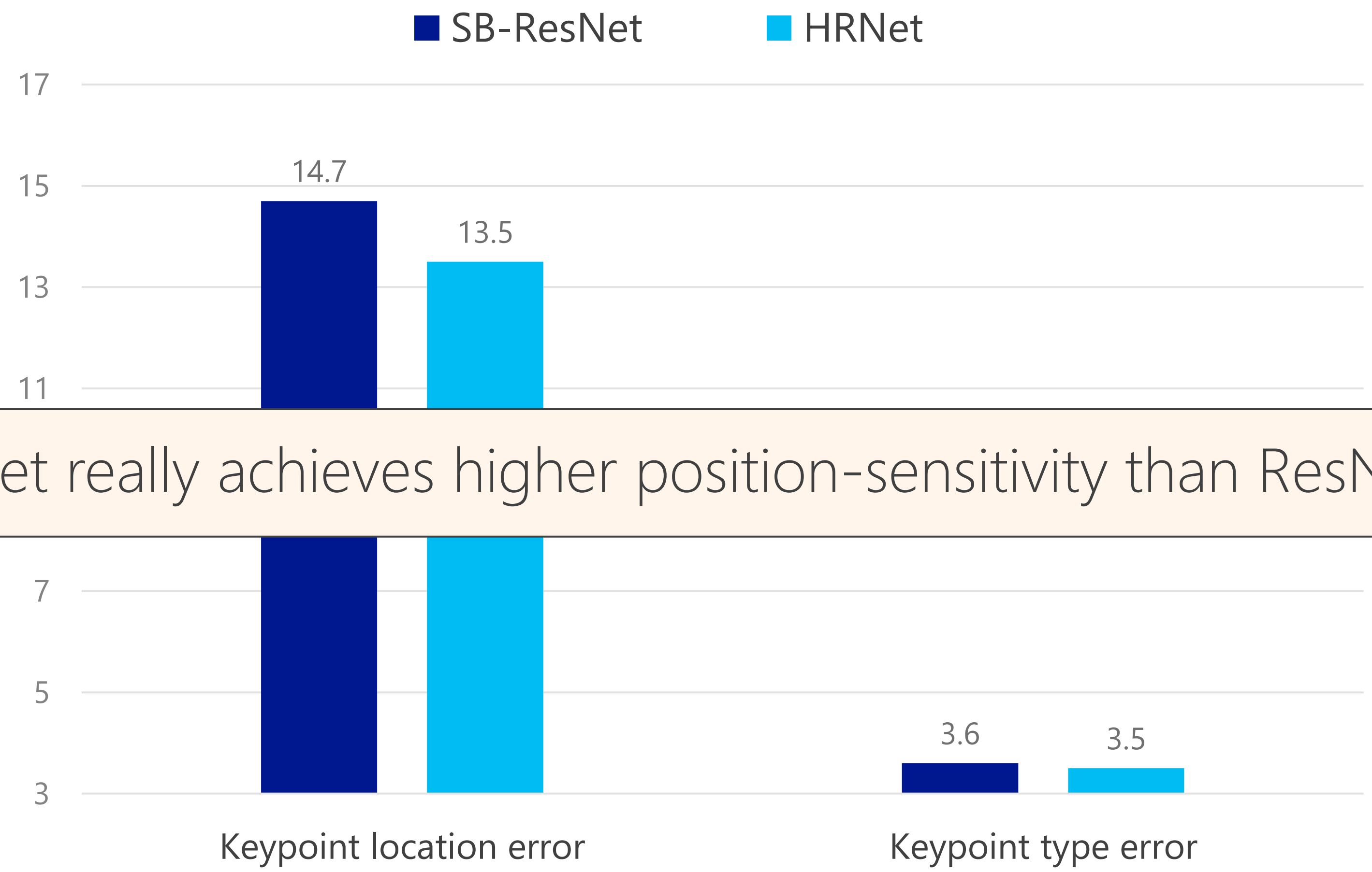
COCO test-dev



COCO test-dev

method	Backbone	Input size	#Params	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
Bottom-up: keypoint detection and grouping										
OpenPose [6], *CMU*	-	-	-	-	61.8	84.9	67.5	57.1	68.2	66.5
Associative Embedding [39]	-	-	-	-	65.5	86.8	72.3	60.6	72.6	70.2
PersonLab [46], *Google*	-	-	-	-	68.7	89.0	75.4	64.1	75.5	75.4
MultiPoseNet [33]	-	-	-	-	69.6	86.3	76.6	65.0	76.3	73.5
Top-down: human detection and single-person keypoint detection										
Mask-RCNN [21], *Facebook*	ResNet-50-FPN	-	-	-	63.1	87.3	68.7	57.8	71.4	-
CPN [11] , *Face++*	ResNet-Inception	384×288	-	-	72.1	91.4	80.0	68.7	77.2	78.5
CPN (ensemble) [11], *Face++*	ResNet-Inception	384×288	-	-	73.0	91.7	80.9	69.5	78.1	79.0
SimpleBaseline [72], *Microsoft*	ResNet-152	384×288	68.6M	35.6	73.7	91.9	81.1	70.3	80.0	79.0
Our approach	HRNet-W32	384×288	28.5M	16.0	74.9	92.5	82.8	71.3	80.9	80.1
Our approach	HRNet-W48	384×288	63.6M	32.9	**75.5**	**92.5**	**83.3**	**71.9**	**81.5**	**80.5**
Our approach + extra data	HRNet-W48	384×288	63.6M	32.9	**77.0**	**92.7**	**84.5**	**73.4**	**83.1**	**82.0**

How does the HRNet improve the quality?

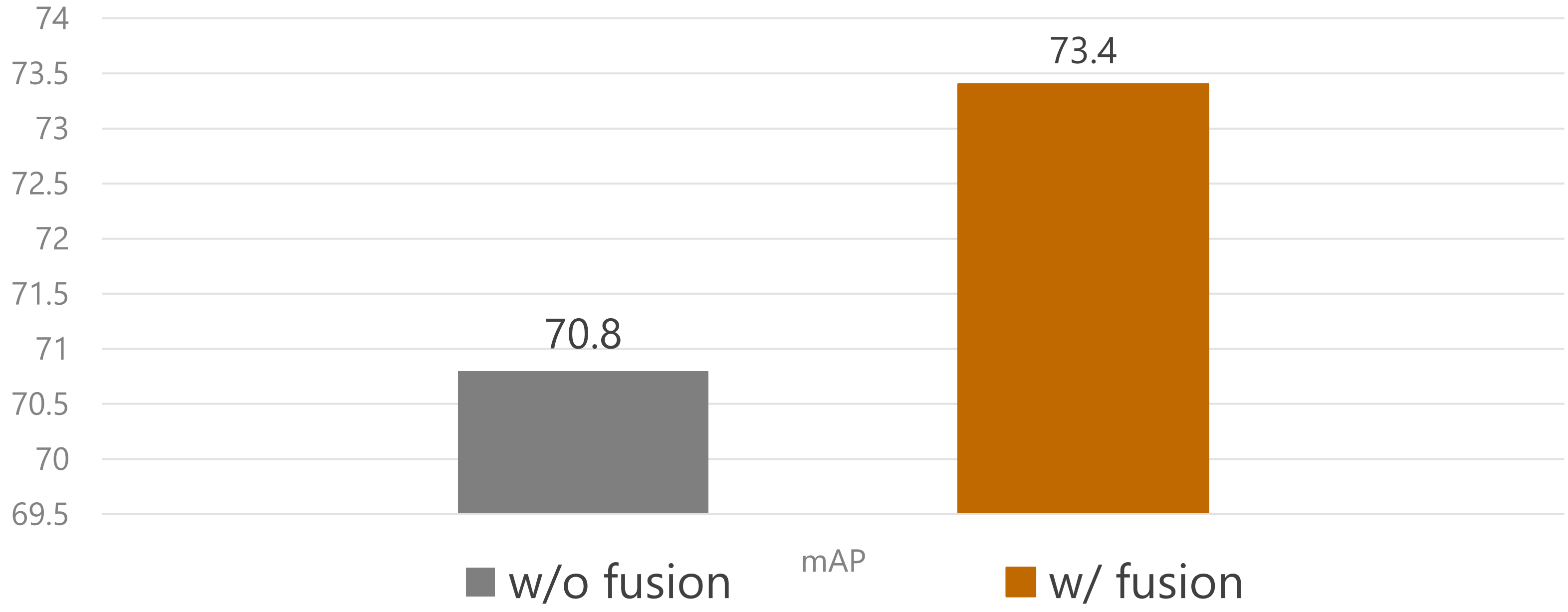


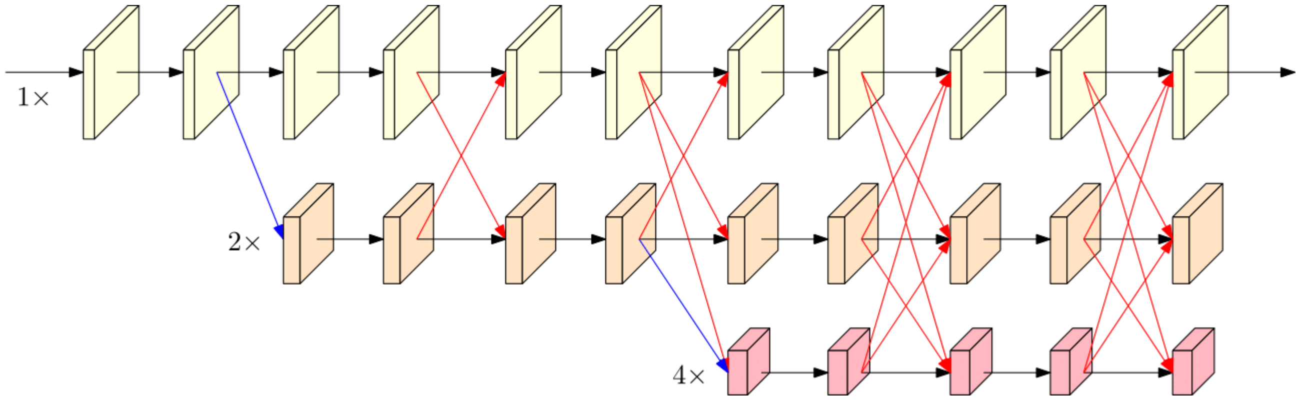
HRNet really achieves higher position-sensitivity than ResNet

Gain = 1.2

Gain = 0.1

Ablation study: repeated across-resolution fusion

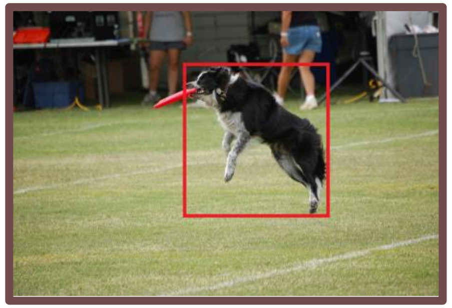




COCO, train from scratch

Visual recognition applications

Image classification



Object detection

Semantic segmentation

Face alignment

Pose estimation

Semantic segmentation

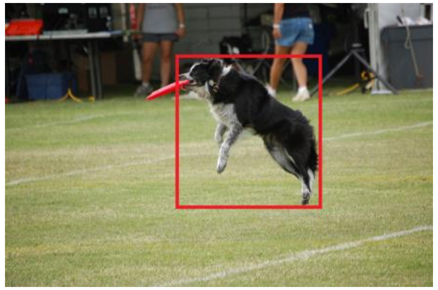


Image classification

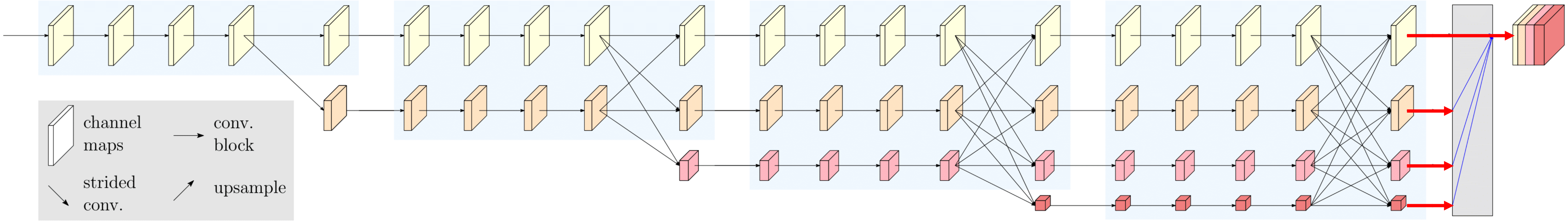
Object detection

Semantic segmentation

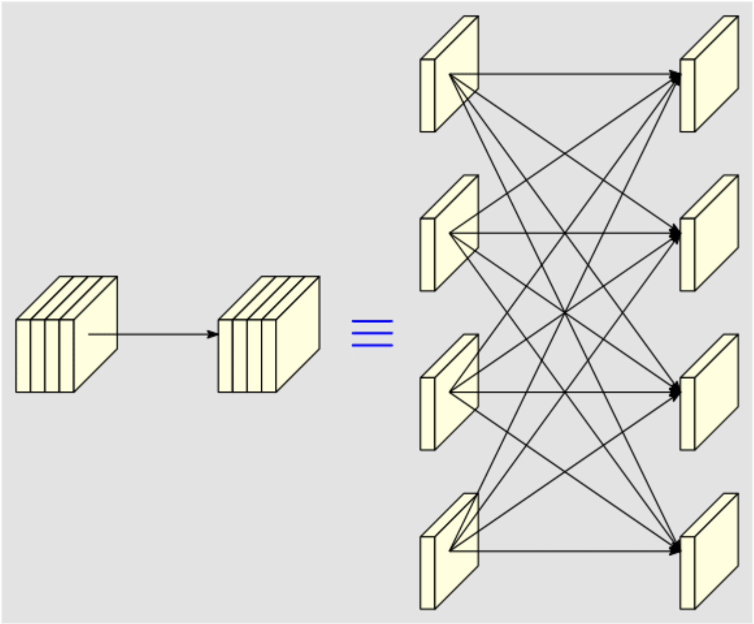
Face alignment

Pose estimation

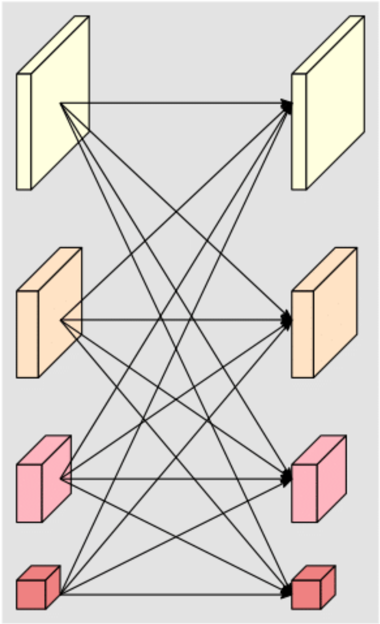
HRNet for segmentation



Relation to regular convolution



Regular convolution



Multi-resolution convolution
(across-resolution fusion)

Cityscapes segmentation validation results

	backbone	#Params.	GFLOPs	mIoU
U-Net++ [130]	ResNet-101	59.5M	748.5	75.5
DeepLabv3 [14], *Google*	Dilated-resNet-101	58.0M	1778.7	78.5
DeepLabv3+ [16], *Google*	Dilted-Xception-71	43.5M	1444.6	79.6
PSPNet [123], *SenseTime*	Dilated-ResNet-101	65.9M	2017.6	79.7
Our approach	HRNet-W40	45.2M	493.2	80.2

Cityscapes segmentation validation results

	backbone	#Params.	GFLOPs	mIoU
U-Net++ [130]	ResNet-101	59.5M	748.5	75.5
DeepLabv3 [14], *Google*	Dilated-resNet-101	58.0M	1778.7	78.5
DeepLabv3+ [16], *Google*	Dilted-Xception-71	43.5M	1444.6	79.6
PSPNet [123], *SenseTime*	Dilated-ResNet-101	65.9M	2017.6	79.7
Our approach	HRNet-W40	45.2M	493.2	80.2
Our approach	HRNet-W48	65.9M	747.3	**81.1**

Cityscapes segmentation testing results

	backbone	mIoU
DeepLab [13], *Google*	Dilated-ResNet-101	70.4
SAC [117]	Dilated-ResNet-101	78.1
DepthSeg [46]	Dilated-ResNet-101	78.2
ResNet38 [101]	WResNet-38	78.4
BiSeNet [111]	ResNet-101	78.9
DFN [112]	ResNet-101	79.3
PSANet [125], *SenseTime*	Dilated-ResNet-101	80.1
PADNet [106]	Dilated-ResNet-101	80.3
DenseASPP [124]	WDenseNet-161	80.6
Our approach	HRNet-W48	**81.6**

Cityscapes segmentation testing results

	backbone	mIoU
DeepLab [13], *Google*	Dilated-ResNet-101	70.4
SAC [117]	Dilated-ResNet-101	78.1
DepthSeg [46]	Dilated-ResNet-101	78.2
ResNet38 [101]	WResNet-38	78.4
BiSeNet [111]	ResNet-101	78.9
DFN [112]	ResNet-101	79.3
PSANet [125], *SenseTime*	Dilated-ResNet-101	80.1
PADNet [106]	Dilated-ResNet-101	80.3
DenseASPP [124]	WDenseNet-161	80.6
Our approach	HRNet-W48	**81.6**
Our approach + OCR	HRNet-W48	**82.3**

Yuhui Yuan, Xilin Chen, Jingdong Wang: Object-Contextual Representations for Semantic Segmentation. CoRR abs/1909.11065 (2019)

PASCAL context

	backbone	mIoU (59classes)	mIoU (60classes)
FCN-8s [86]	VGG-16	-	35.1
BoxSup [20]	-	-	40.5
HO_CRF [1]	-	-	41.3
Piecewise [60]	VGG-16	-	43.3
DeepLabv2 [13], *Google*	Dilated-ResNet-101	-	45.7
RefineNet [59]	ResNet-152	-	47.3
U-Net++ [130]	ResNet-101	47.7	-
PSPNet [123], *SenseTime*	Dilated-ResNet-101	47.8	-
Ding et al. [23]	ResNet-101	51.6	-
EncNet [114]	Dilated-ResNet-101	52.6	-
Our approach	HRNetV2-W48	**54.0**	**48.3**
Our approach + OCR	HRNetV2-W48	**56.2**	**-**

LIP validation

	backbone	extra	pixel acc.	avg. acc.	mIoU
Attention+SSL [34]	VGG-16	Pose	84.36	54.94	44.73
DeepLabv2 [16], *Google*	Dilated-ResNet-101	-	84.09	55.62	44.80
MMAN[67]	Dilated-ResNet-101	-	-	-	46.81
SS-NAN [125]	ResNet-101	Pose	87.59	56.03	47.92
MuLA [72]	Hourglass	Pose	88.50	60.50	49.30
JPPNet [57]	Dilated-ResNet-101	Pose	86.39	62.32	51.37
CE2P [65]	Dilated-ResNet-101	Edge	87.37	63.20	53.10
Our approach	HRNetV2-W48	N	**88.21**	**67.43**	**55.90**
Our approach + OCR	HRNetV2-W48	N	**-**		**56.66**

Object detection

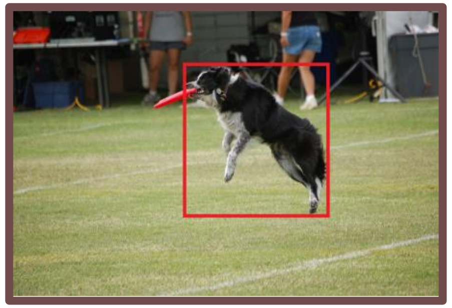


Image classification

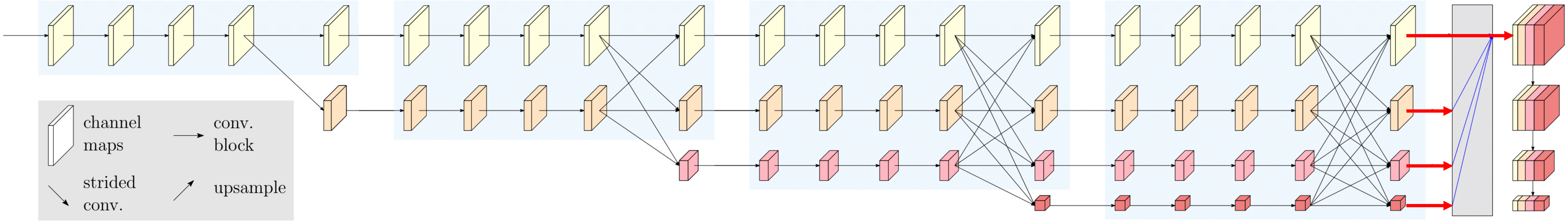
Object detection

Semantic segmentation

Face alignment

Pose estimation

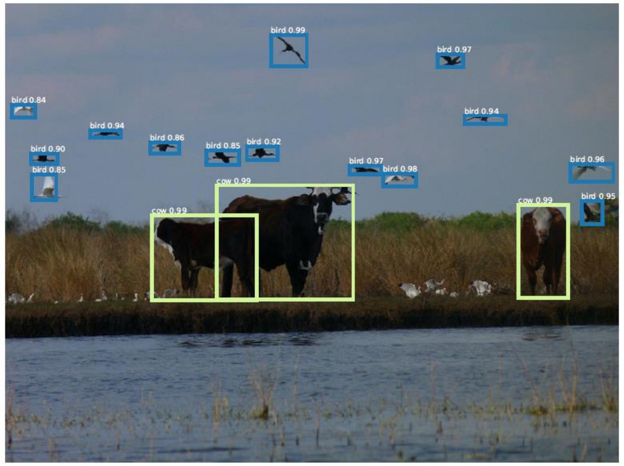
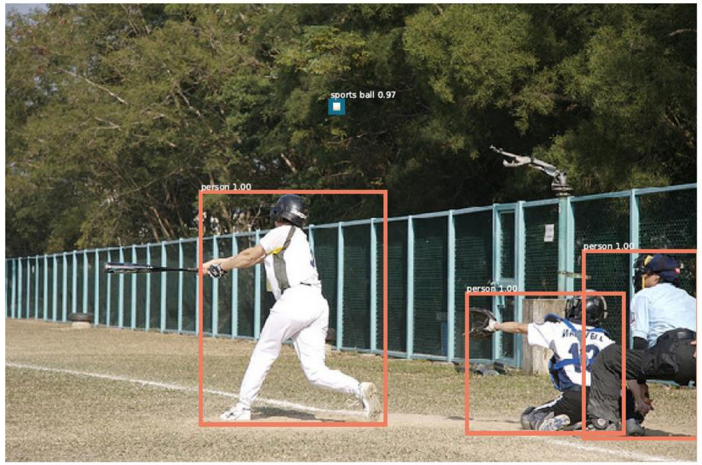
HRNet-FPN for object detection

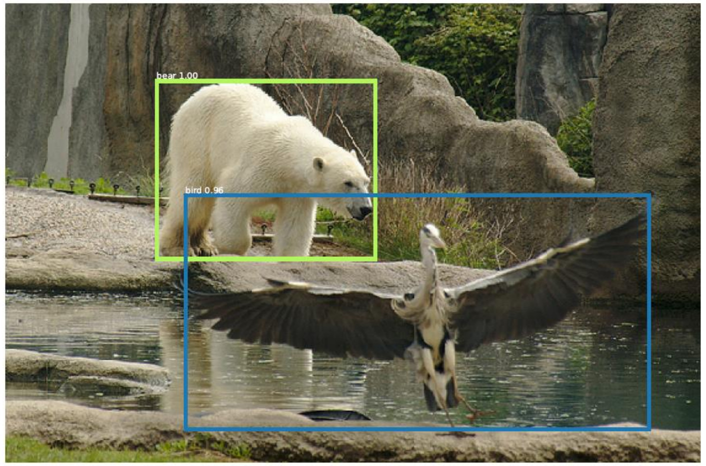


Faster R-CNN

	Backbone	Size	LS	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Faster R-CNN [61]	ResNet-101-FPN	800	$2\times$	40.3	61.8	43.9	22.6	43.1	51.0
Faster R-CNN	HRNet-W32-FPN	800	$2\times$	41.1	62.3	44.9	24.0	43.1	51.4
Faster R-CNN [61]	ResNet-152-FPN	800	$2\times$	40.6	62.1	44.3	22.6	43.4	52.0
Faster R-CNN	HRNet-W40-FPN	800	$2\times$	42.1	63.2	46.1	24.6	44.5	52.6
Faster R-CNN [11]	ResNeXt-101-64x4d-FPN	800	$2\times$	41.1	62.8	44.8	23.5	44.1	52.3
Faster R-CNN	HRNet-W48-FPN	800	$2\times$	42.4	**63.6**	46.4	24.9	44.6	53.0
Cascade R-CNN [9]	ResNet-101-FPN	800	$\sim 1.6\times$	42.8	62.1	46.3	23.7	45.5	55.2
Cascade R-CNN	HRNet-W32-FPN	800	$\sim 1.6\times$	**43.7**	62.0	**47.4**	**25.5**	46.0	**55.3**

single model single scale

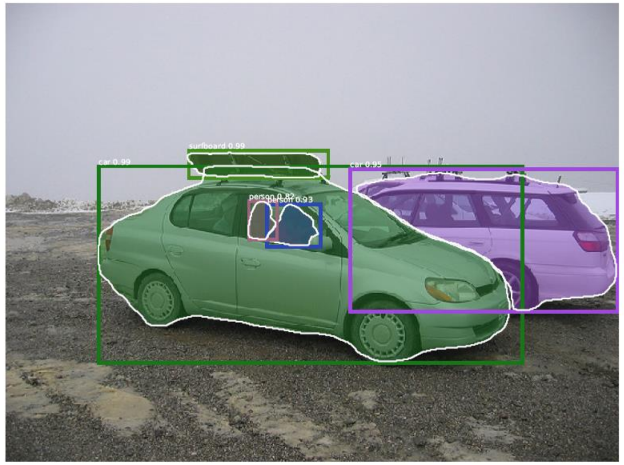
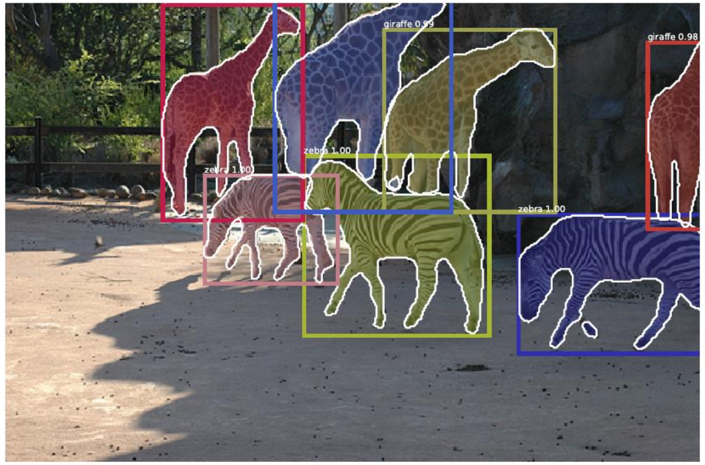





Mask R-CNN

backbone	LS	mask				bbox			
		AP	AP_S	AP_M	AP_L	AP	AP_S	AP_M	AP_L
ResNet-50-FPN	2 ×	35.0	16.0	**37.5**	**52.0**	38.6	21.7	41.6	50.9
HRNet-W18-FPN	2 ×	**35.3**	**16.9**	**37.5**	51.8	**39.2**	**23.7**	**41.7**	**51.0**
ResNet-101-FPN	2 ×	36.7	17.0	39.5	54.8	41.0	23.4	44.4	53.9
HRNet-W32-FPN	2 ×	**37.6**	**17.8**	**40.0**	**55.0**	**42.3**	**25.0**	**45.4**	**54.9**

single model single scale


In addition, we obtain better detection/instance segmentation results under the very recent frameworks: FCOS, CenterNet, and Hybrid Task Cascade

Image classification

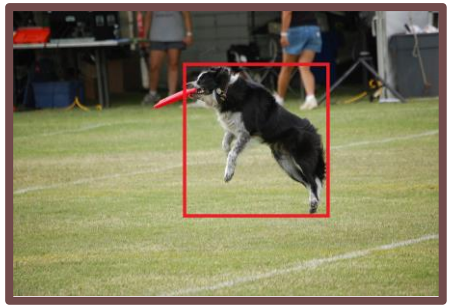


Image classification

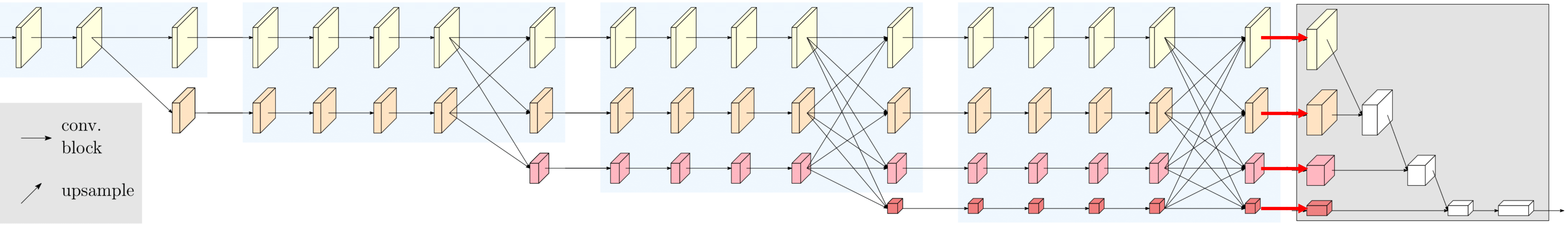
Object detection

Semantic segmentation

Face alignment

Pose estimation

HRNet for ImageNet classification



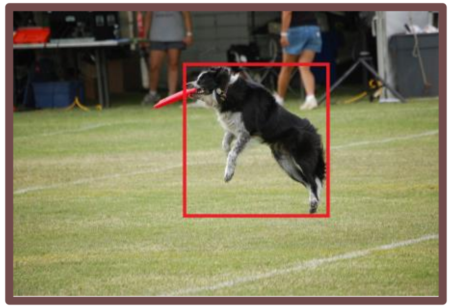
ImageNet classification results

	#Params.	GFLOPs	Top-1 err.	Top-5 err.
ResNet-50	25.6M	3.82	23.3%	6.6%
HRNet-W44	21.9M	3.90	**23.0%**	**6.5%**
ResNet-101	44.6M	7.30	21.6%	5.8%
HRNet-W76	40.8M	7.30	**21.5%**	**5.8%**
ResNet-152	60.2M	10.7	21.2%	5.7%
HRNet-W96	57.5M	10.2	**21.0%**	**5.7%**

HRNet performs slightly better than ResNet

HRNet applications

Image classification



Object detection

Semantic segmentation

Face alignment

Pose estimation

Discussions

high-resolution networks
vs
classification networks

neural architecture design (NAD)
vs
neural architecture search (NAS)

HRNet vs classification network

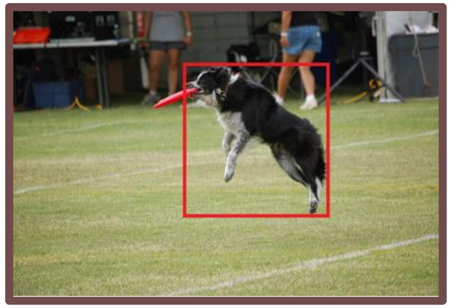




image-level | region-level | pixel-level

Low resolution
High resolution

Recover from low-resolution (ResNet, VGGNet) ✗
High-resolution (our HRNet) ✓

Discussions

high-resolution networks
vs
classification networks

neural architecture design (NAD)
vs
neural architecture search (NAS)

NAD expands search space for NAS

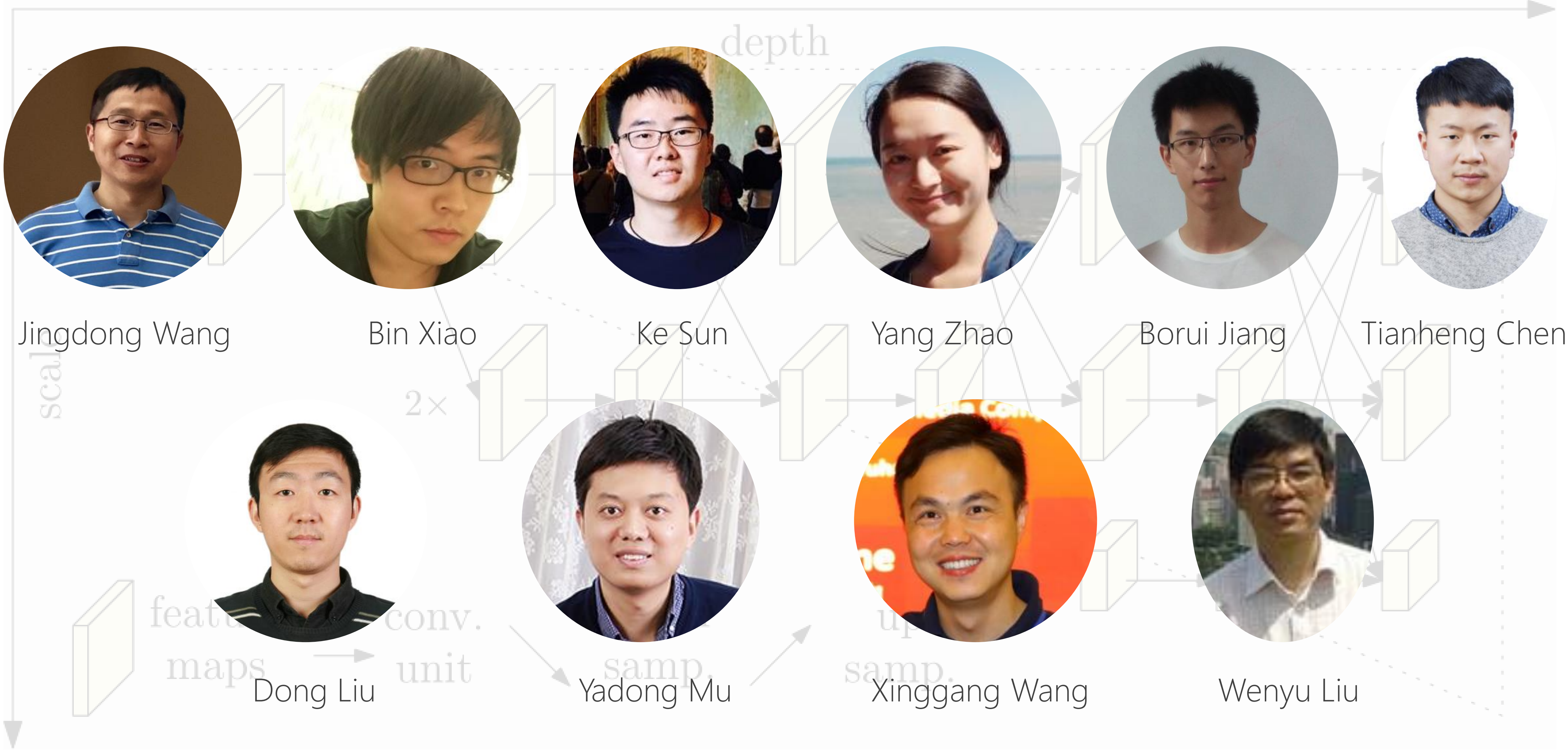
HRNet has been widely used

- The top 5 winners at tiger pose estimation challenge adopted HRNet
- Almost all the COCO keypoint and densepose, mapillary panoptic segmentation winners (joint COCO and Mapillary Recognition workshop, ICCV 2019) adopted HRNet. The modified HRNet achieves the SOTA performance on mapillary panoptic segmentation for a single model
- The winner in CVPR 2019 image enhancement challenge adopted HRNet
- The winner in CVPR 2019 LIP pose estimation challenge adopted HRNet
- HRNet is combined into the MMDetection framework: superior object detection and instance segmentation over ResNet and ResNeXt
- The AzureCAT team adopted HRNet for satellite and seismic image parsing
- Lane line detection, long distance car detection for auto-driving
- Image translation, stylization
- ……

Conclusions

- **Design from scratch and maintain** high-resolution representations through the whole process with repeated across-resolution fusions.
- **Fundamental architecture change**. Different from the previous standard design (connect high-to-low convolutions in series) that originates from LeNet-5 by Yann LeCun
- **A generic network.** Capable of learning strong high-resolution representations. and superior in many position-sensitive vision tasks than ResNet and VGGNet: semantic segmentation, object detection, facial landmark detection, human pose estimation, salient object detection, edge detection, and image-to-image translation, image stylization …

HRNet team


Jingdong Wang
Bin Xiao
Ke Sun
Yang Zhao
Borui Jiang
Tianheng Cheng
Dong Liu
Yadong Mu
Xinggang Wang
Wenyu Liu

Thanks!
Q&A

Human pose estimation

Segmentation, detection, alignment, classification