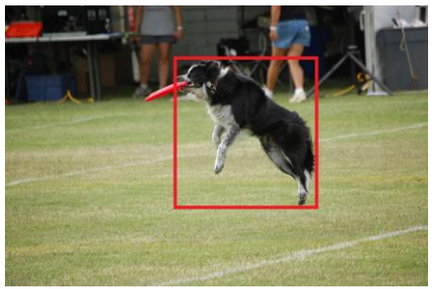
High-Resolution Networks

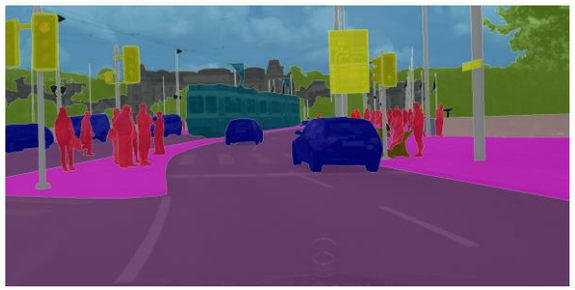
Jingdong Wang
Senior Principal Research Manager
Microsoft Research, Beijing, China

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

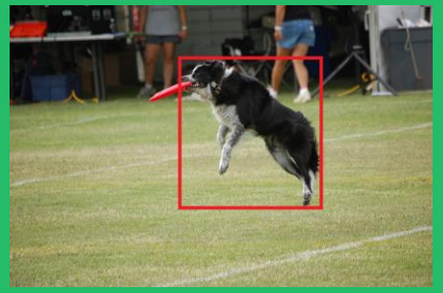


image classification

region-level recog.

pixel-level recog.

global

position-sensitive

Low-resolution representation learning

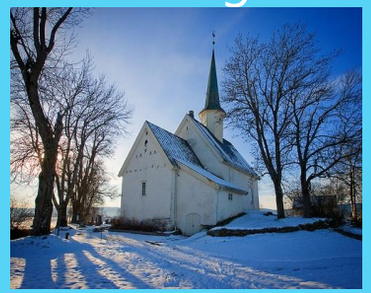
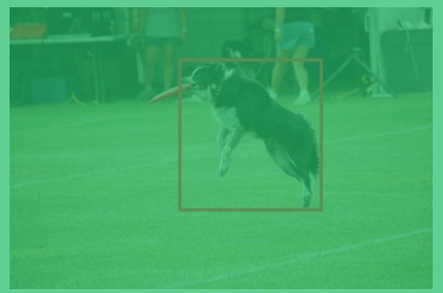


image classification

global

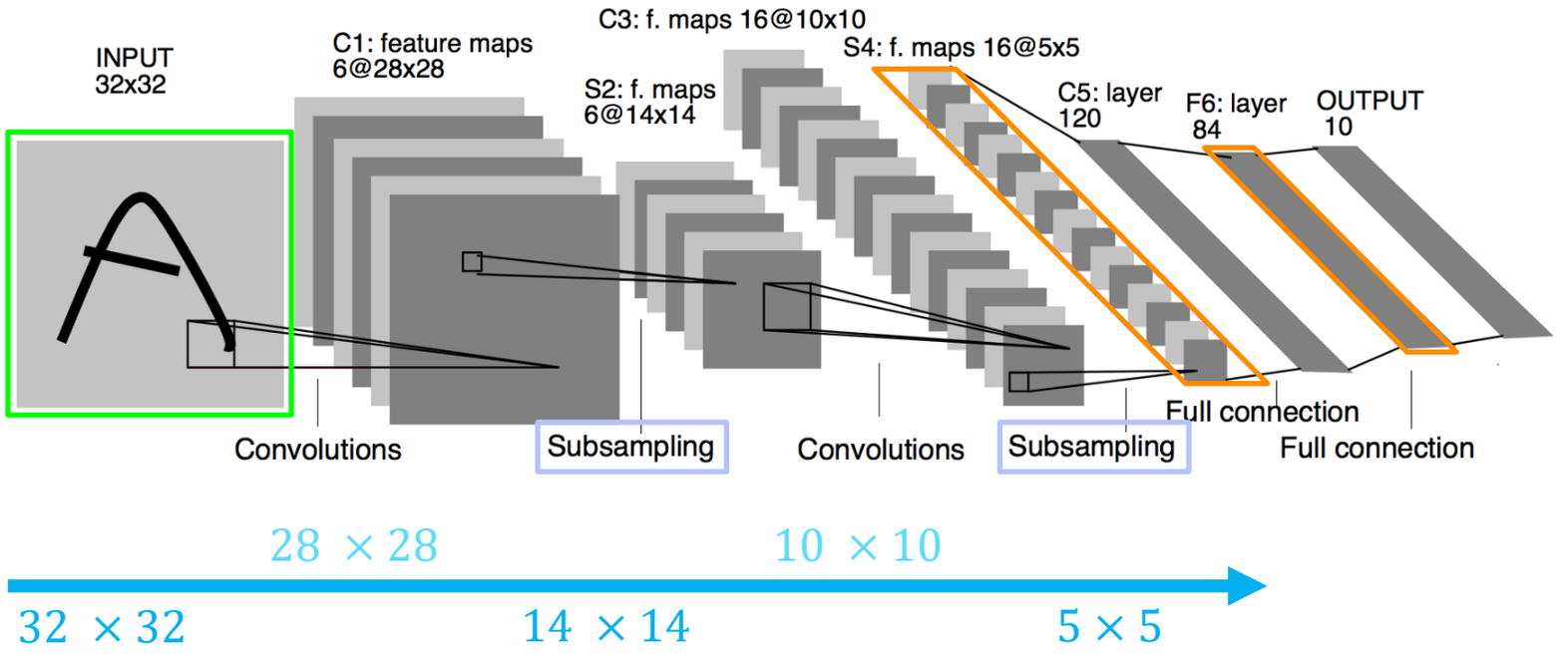
High-resolution



Low-resolution representation learning

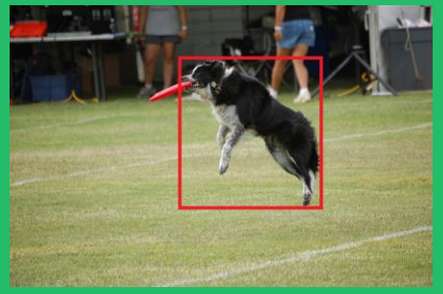
Classification networks: connect the convolutions in *series* from high resolution to low resolution

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



LeNet-5 (Yann LeCun,1998)

High-resolution representation learning

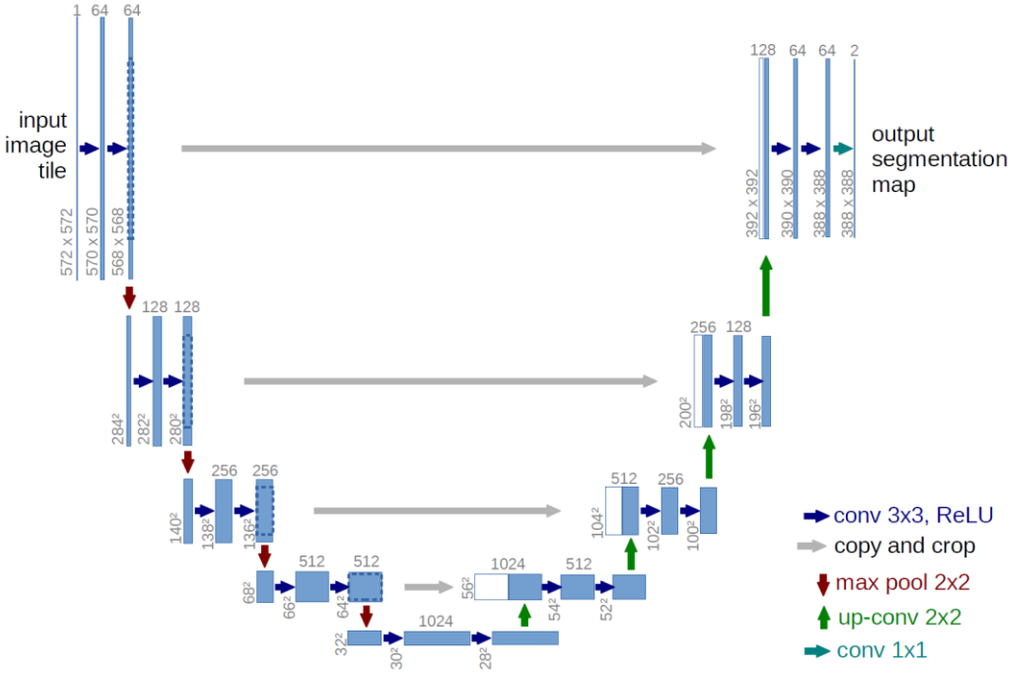


region-level recog.

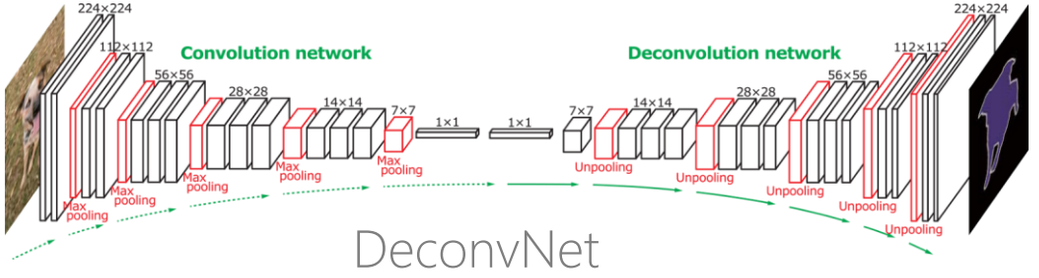
pixel-level recog.

position-sensitive

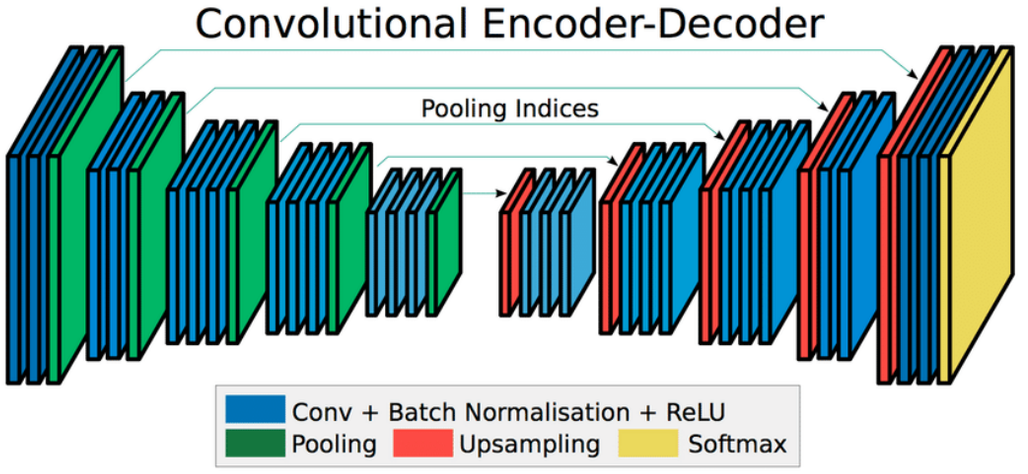
Previous high-resolution representation learning



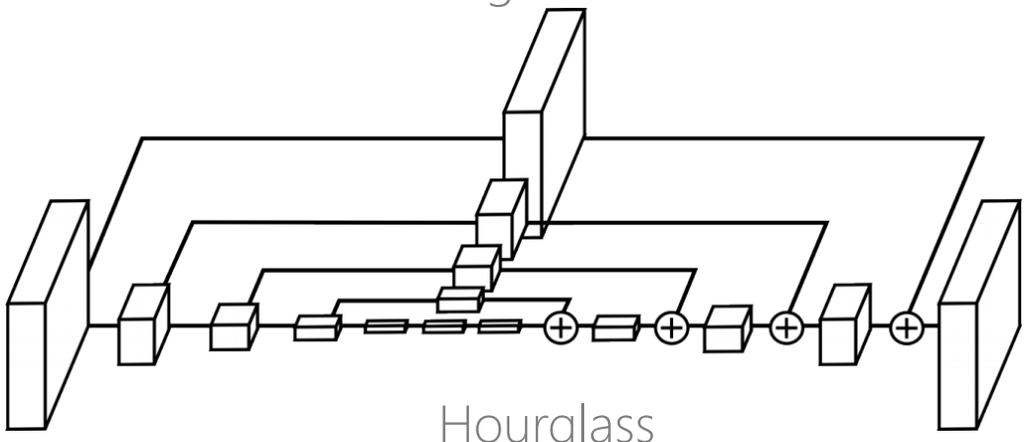
U-Net



DeconvNet



SegNet



Hourglass

Previous SOTA solutions: look different, essentially the same

Previous high-resolution representation learning

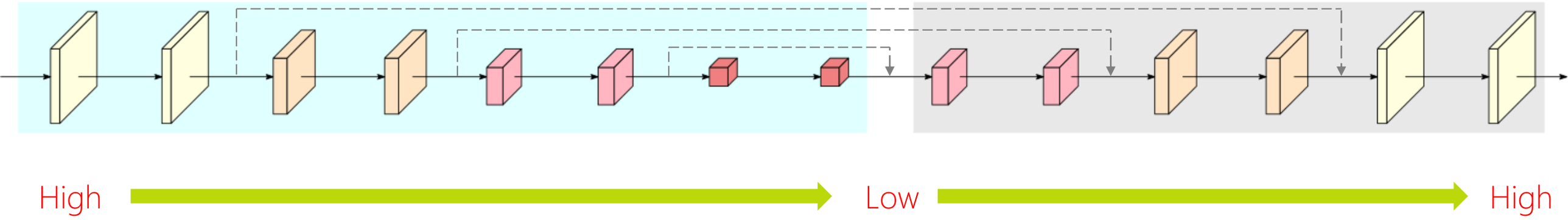
Essentially, the previous methods **remediate/extend** classification networks

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

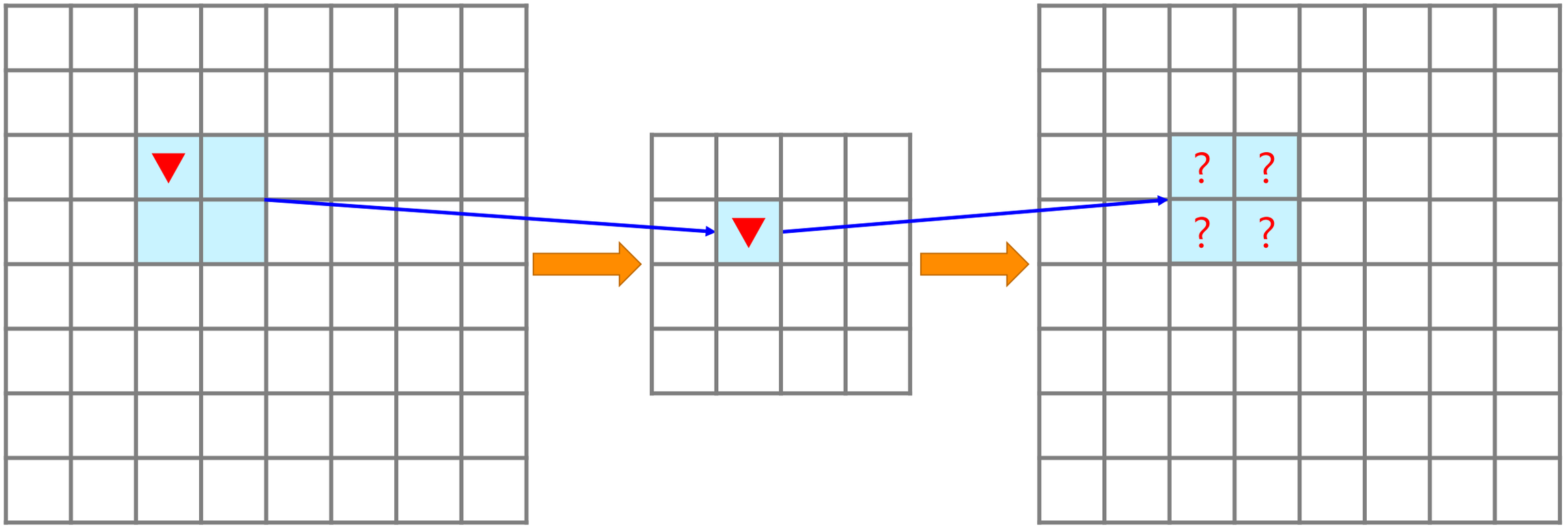
Previous high-resolution representation learning

Essentially, the previous methods **remediate/extend** classification networks

- **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)

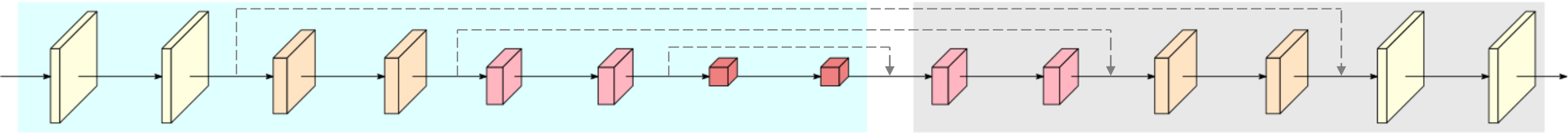
8×8 (high)

Previous high-resolution representation learning

Essentially, the previous methods **remediate/extend** classification networks

- **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



High → Low → High

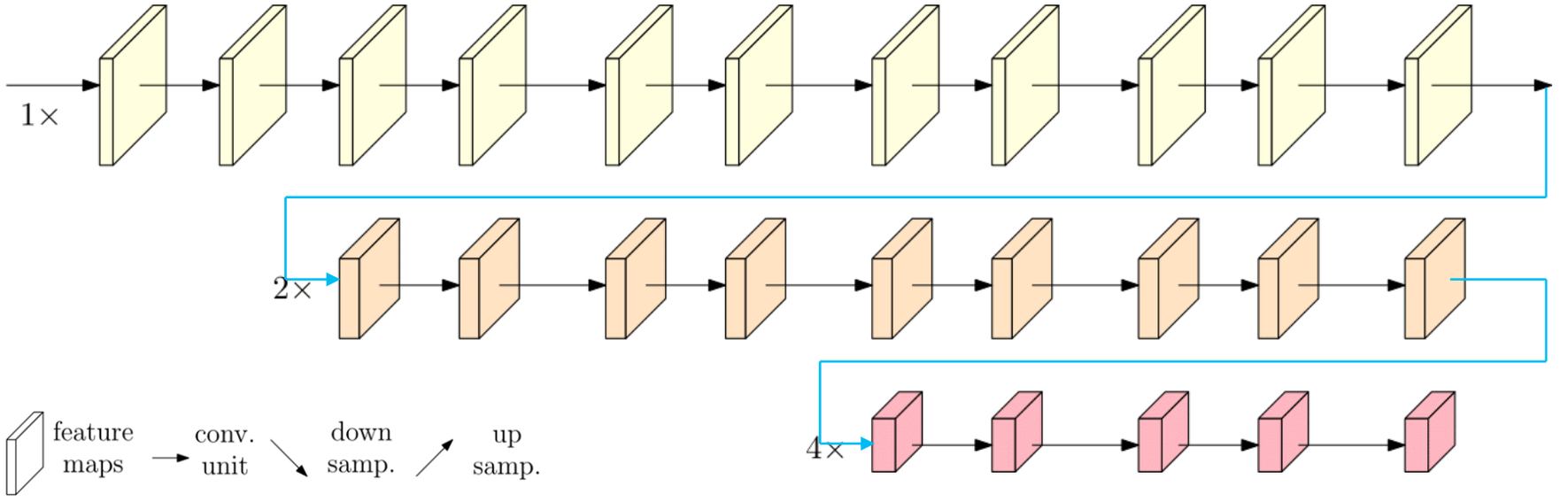
Our work: High-resolution network (HRNet)

- 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 (Accepted by TPAMI)

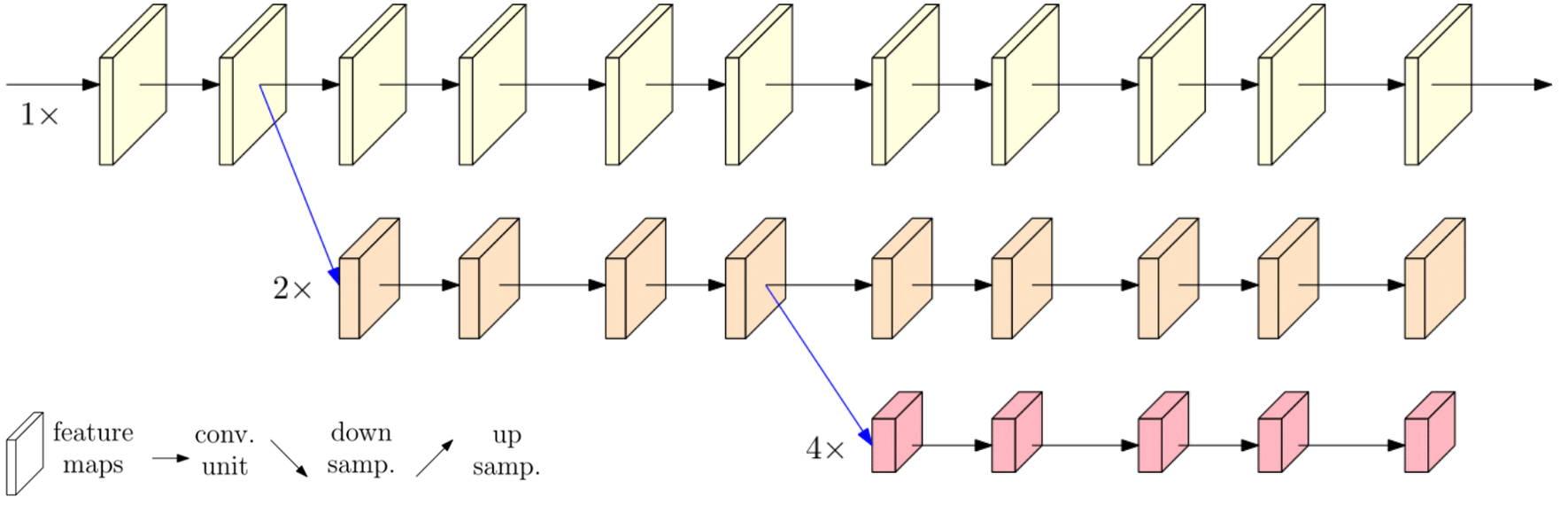
Previous low-resolution networks

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



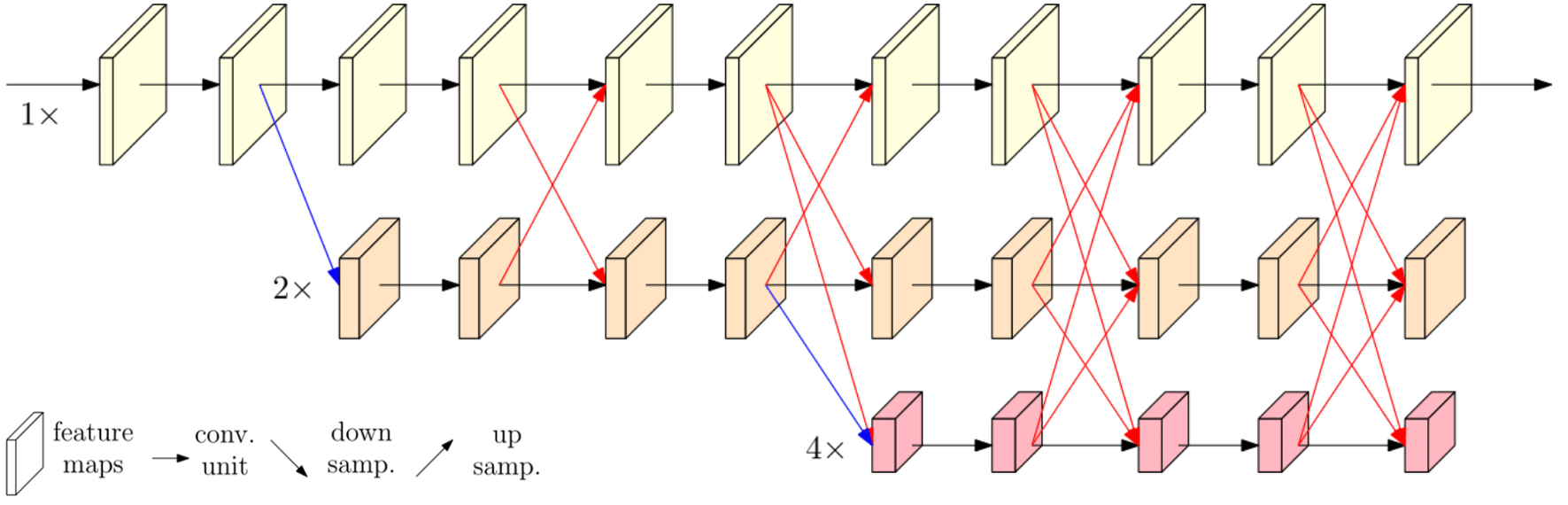
HRNet: high-resolution representation learning

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



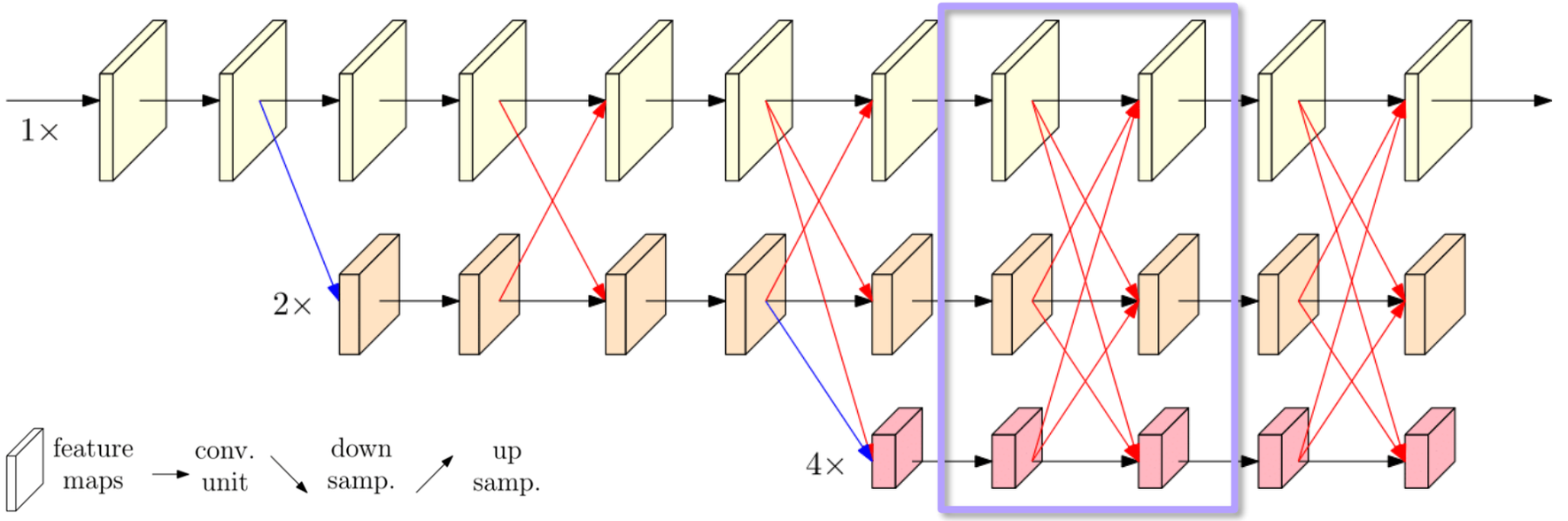
HRNet: high-resolution representation learning

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

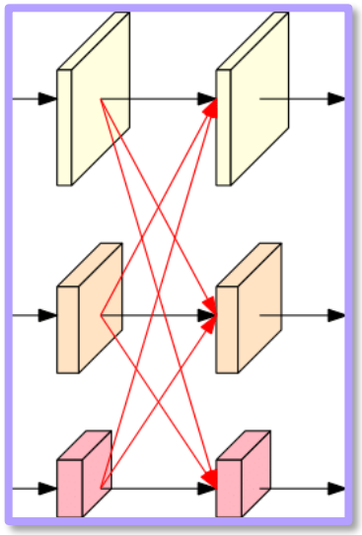


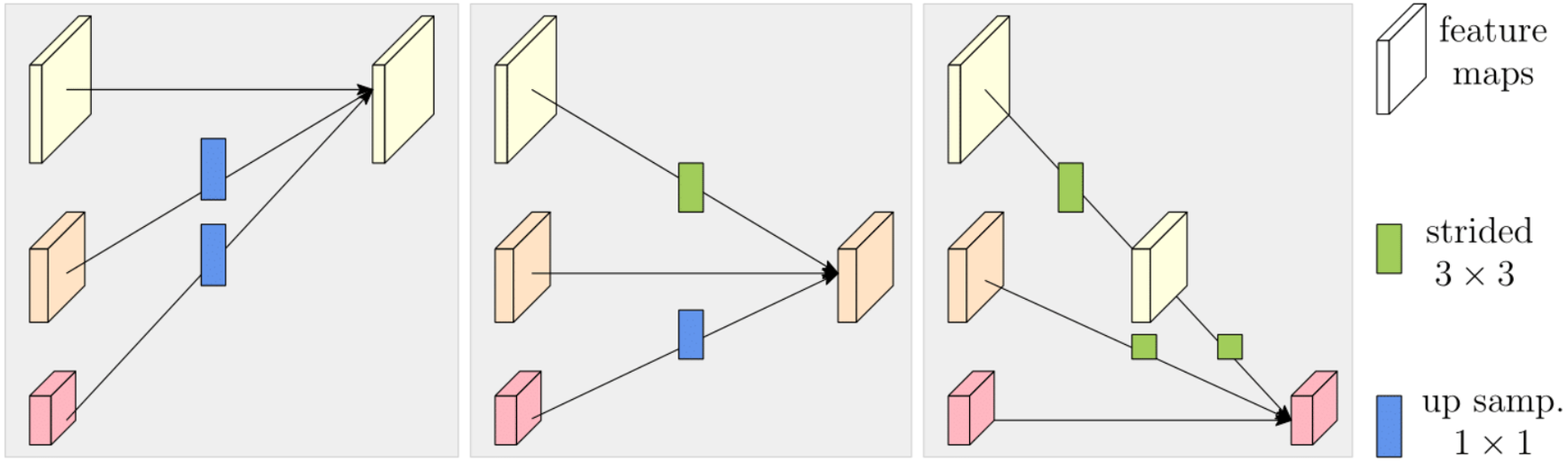
HRNet: high-resolution representation learning

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



Across-resolution fusion





Down-sample: $\mathbf{stride-2}\ 3 \times 3$

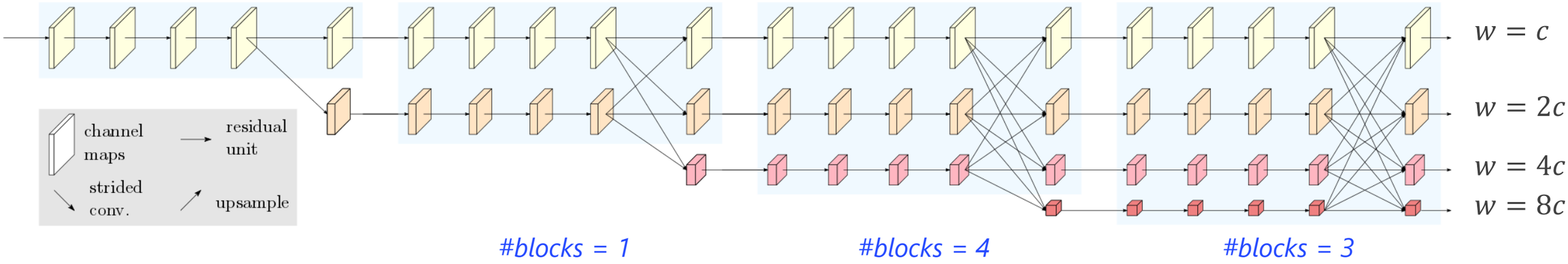
Up-sample: $\mathbf{bilinear} + 1 \times 1$

Fundamental architecture changes

- Connect high-to-low resolution convolutions in ~~*series*~~ parallel
- ~~*Recover*~~ Maintain high-resolution representations ~~*from low-resolution representations*~~ through the whole process
- Repeat fusions across resolutions to strengthen high- & low-resolution representations

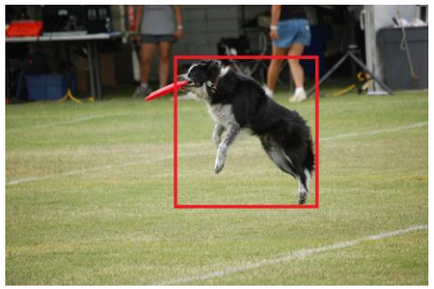
HRNet can learn *high-resolution* representations with *strong position sensitivity*

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

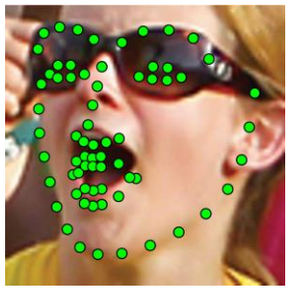



Image classification

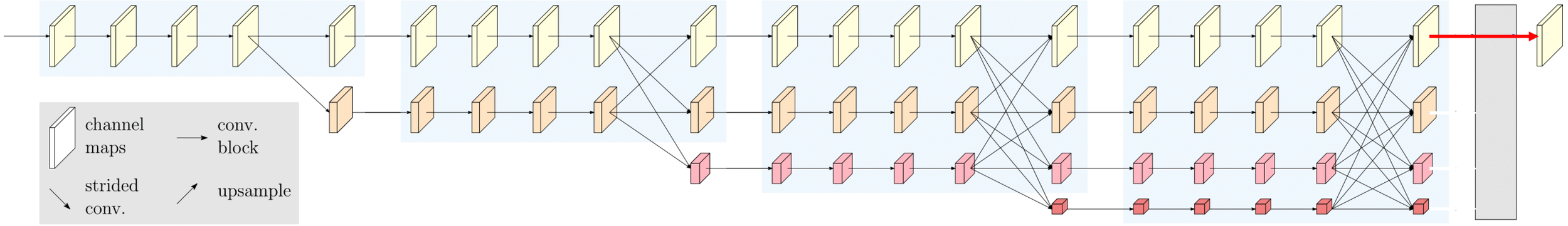
Object detection

Semantic segmentation

Face alignment

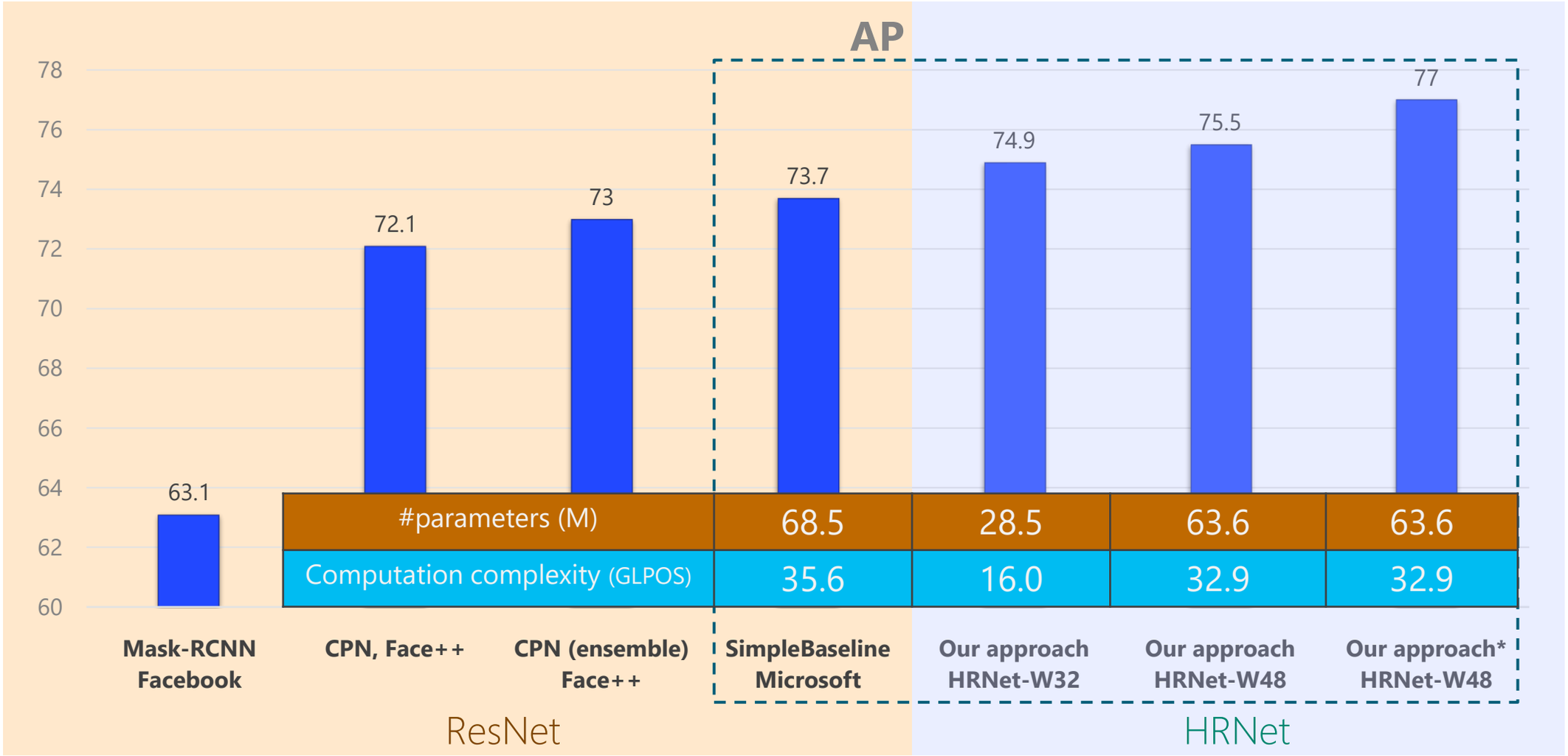
Pose estimation

HRNet for human pose estimation



COCO human pose estimation

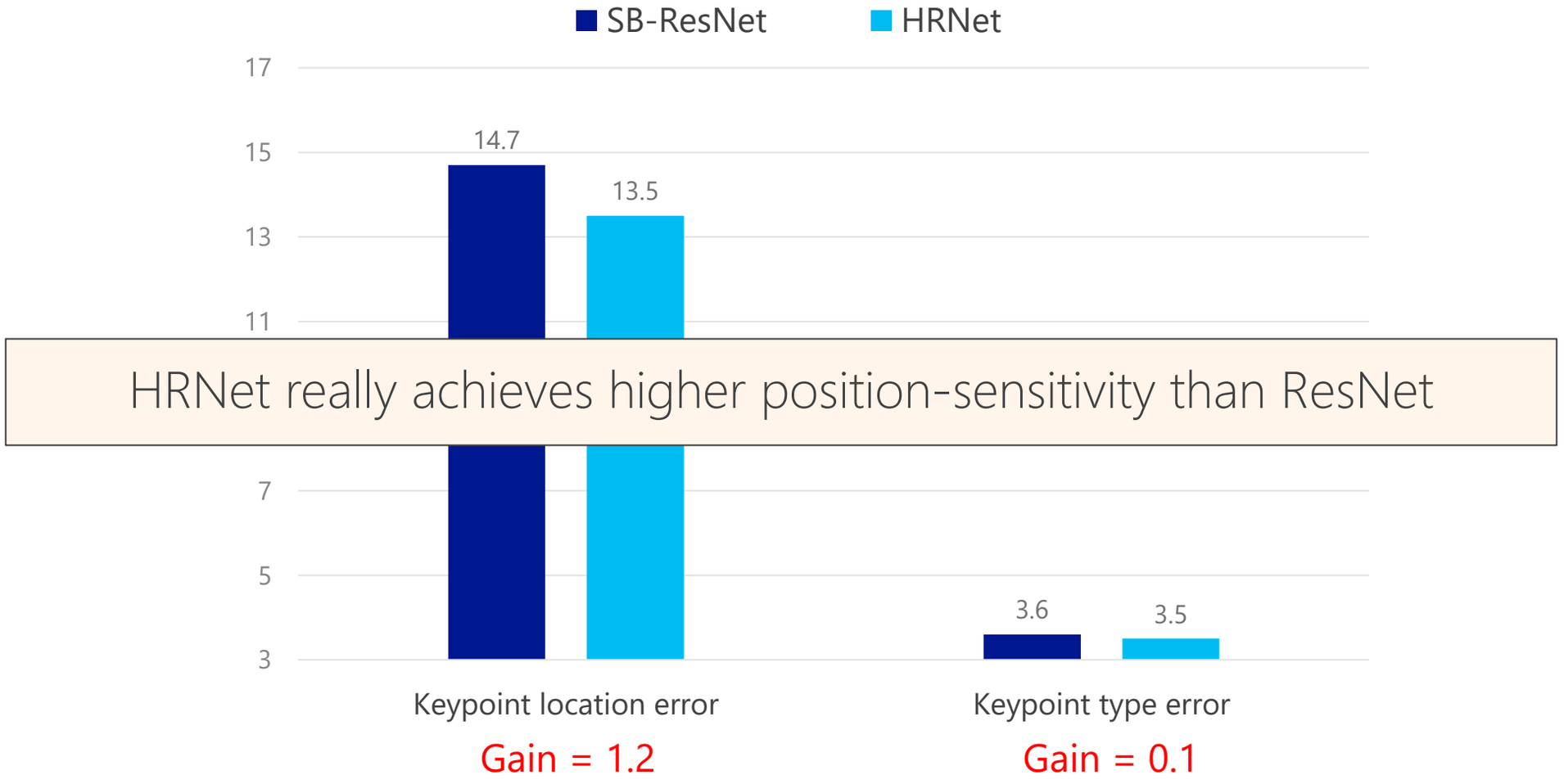
COCO test-dev



COCO test-dev

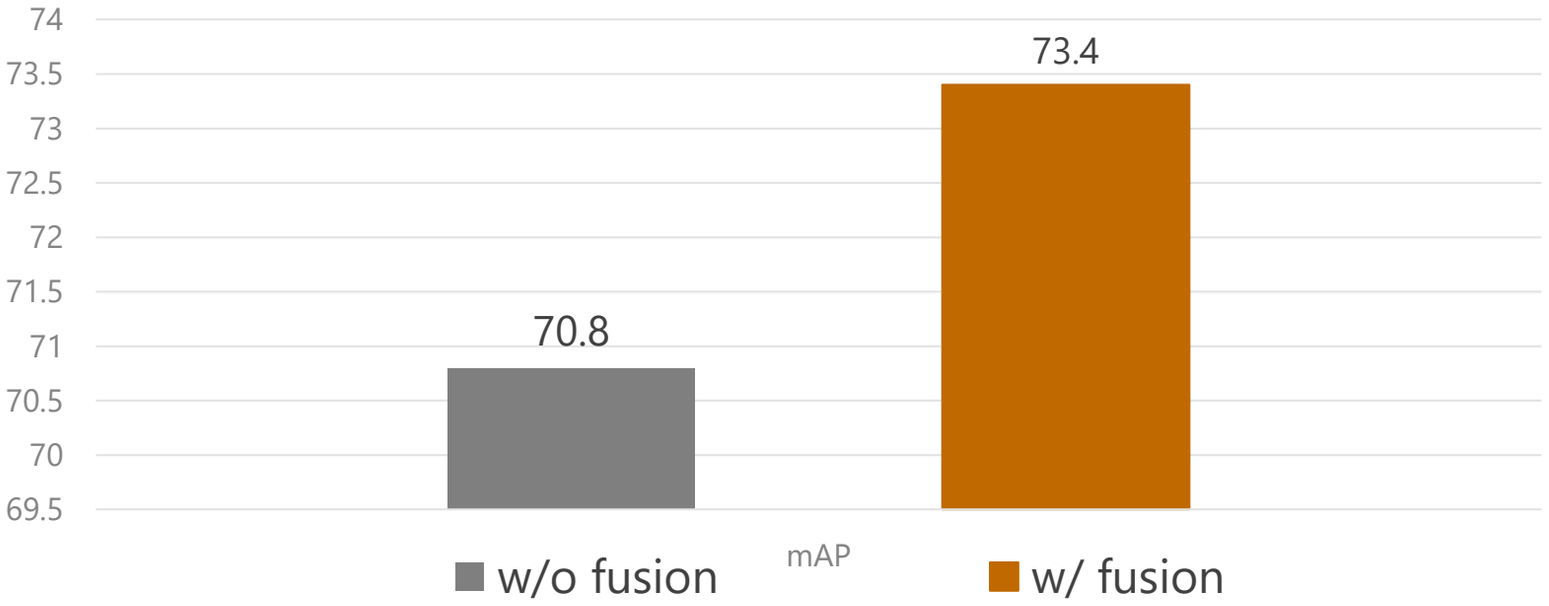
method	Backbone	Input size	#Params	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
Bottom-up: keypoint detection and grouping										
OpenPose [6], *CMU*	-	-	-	-	61.8	84.9	67.5	57.1	68.2	66.5
Associative Embedding [39]	-	-	-	-	65.5	86.8	72.3	60.6	72.6	70.2
PersonLab [46], *Google*	-	-	-	-	68.7	89.0	75.4	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**

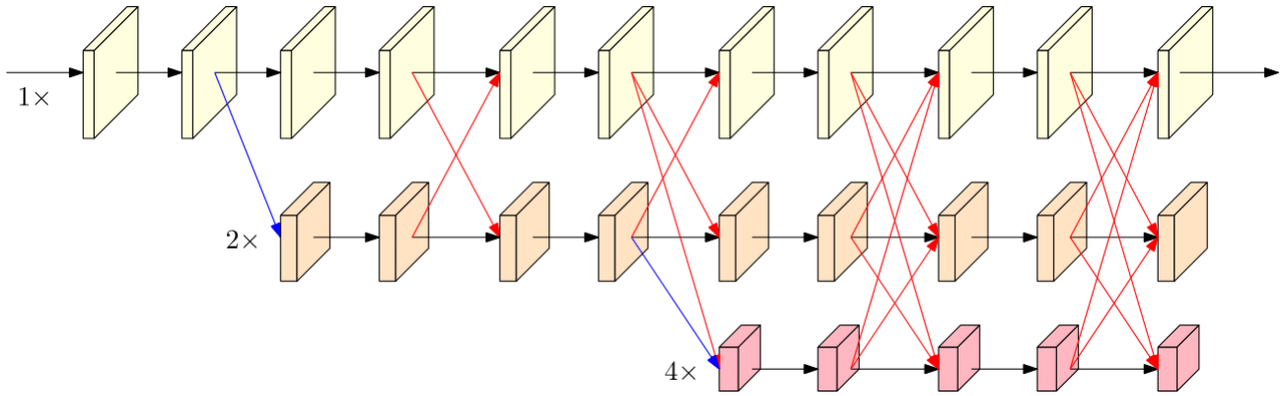
How does the HRNet improve the quality?



HRNet really achieves higher position-sensitivity than ResNet

Ablation study: repeated across-resolution fusion





COCO, train from scratch

Visual recognition applications



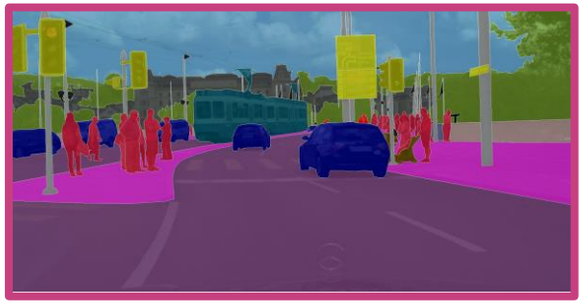


Image classification

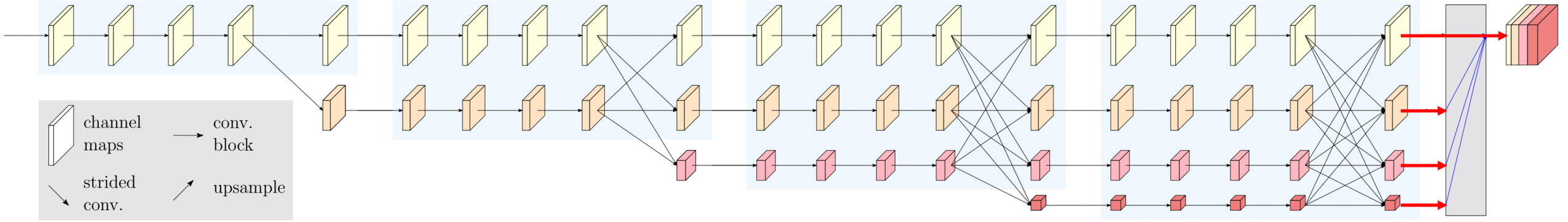
Object detection

Semantic segmentation

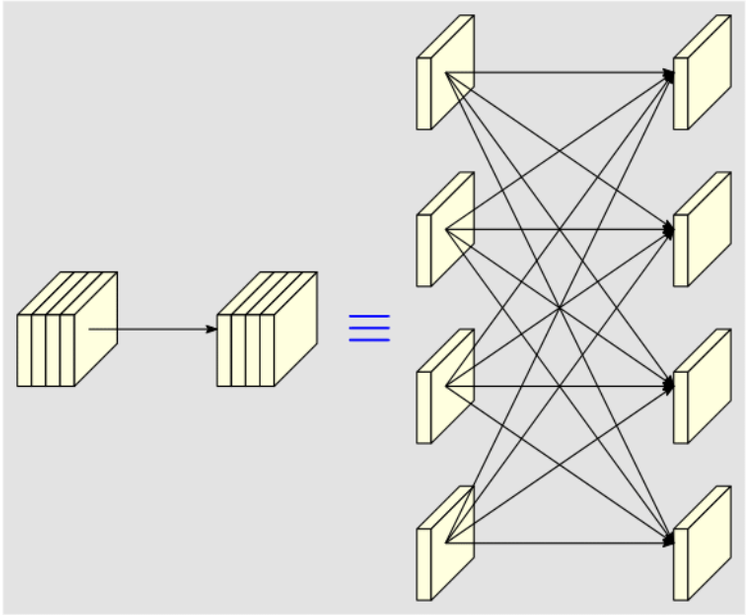
Face alignment

Pose estimation

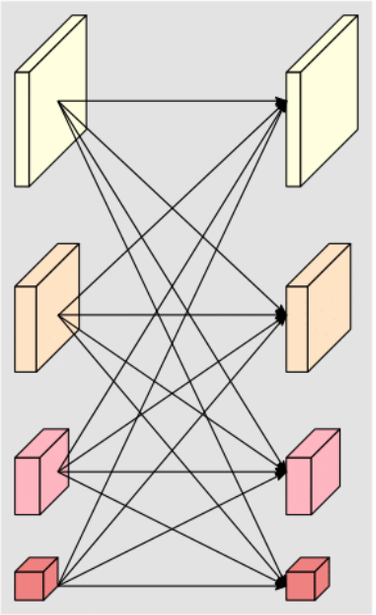
HRNet for face alignment



Relation to regular convolution



Regular convolution



Multi-resolution convolution
(across-resolution fusion)

Datasets and evaluation

Datasets	training	testing	#landmarks	Evaluation (NME)
WFLW	7500	2500	98	inter-ocular
AFLW	20000	full 4386, frontal 1314	19	box
COFW	1345	507	29	inter-ocular
300W	3148	full 689, challenging 135, test 600	68	inter-ocular

WFLW

	backbone	Test	Pose	Expr	Illu	Mu	Occu	Blur
ESR [10]	-	11.13	25.88	11.47	10.49	11.05	13.75	12.20
SDM [105], *CMU*	-	10.29	24.10	11.45	9.32	9.38	13.03	11.28
CFSS [132]	-	9.07	21.36	10.09	8.30	8.74	11.76	9.96
DVLN [100]	VGG-16	6.08	11.54	6.78	5.73	5.98	7.33	6.88
Our approach	HRNet-W18	**4.60**	**7.94**	**4.85**	**4.55**	**4.29**	**5.44**	**5.42**

AFLW

	backbone	full	frontal
RCN [39]	-	5.60	5.36
CDM [113]	-	5.43	3.77
ERT [43]	-	4.35	2.75
LBF [79]	-	4.25	2.74
SDM [105], *CMU*	-	4.05	2.94
CFSS [131]	-	3.92	2.68
RCPR [8]	-	3.73	2.87
CCL [132]	-	2.72	2.17
DAC-CSR [29]	-	2.27	1.81
TSR [68]	VGG-S	2.17	-
CPM + SBR [25]	CPM	2.14	-
SAN [24]	ResNet-152	1.91	1.85
DSRN [69]	-	1.86	-
LAB (w/o B) [99], *SenseTime*	Hourglass	1.85	1.62
Our approach	HRNet-W18	**1.57**	**1.46**

COFW

	backbone	NME	FR0.1
Human	-	5.60	-
ESR [10]	-	11.20	36.00
RCPR [8]	-	8.50	20.00
HPM [32]		7.50	13.00
CCR [27]	-	7.03	10.90
DRDA [115]	-	6.46	6.00
RAR [102]	-	6.03	4.14
DAC-CSR [29]	-	6.03	4.73
LAB (w/o B) [99], *SenseTime*	Hourglass	5.58	2.76
Our approach	HRNetV2-W18	**3.45**	**0.19**

300W

	backbone	common	challenging	full
RCN [39]	-	4.67	8.44	5.41
DSRN [69]	-	4.12	9.68	5.21
PCD-CNN [50]	-	3.67	7.62	4.44
CPM + SBR [25]	CPM	3.28	7.58	4.10
SAN [24]	ResNet-152	3.34	6.60	3.98
DAN [49]	-	3.19	5.24	3.59
Our approach	HRNetV2-W18	**2.87**	**5.15**	**3.32**

300W test

	backbone	NME	AUC0.08	AUC0.1	FR0.08	FR0.1
Balt. et al. [3]	-	-	19.55	-	38.83	-
ESR [10]	-	8.47	26.09	-	30.50	-
ERT [43]	-	8.41	27.01	-	28.83	-
LBF [79]	-	8.57	25.27	-	33.67	-
Face++ [127]	-	-	32.81	-	13.00	-
SDM [105], *CMU*	-	5.83	36.27	-	13.00	-
CFAN [116]	-	5.78	34.78	-	14.00	-
Yan et al. [108]	-	5.74	34.97	-	12.67	-
CFSS [131]	-	4.78	36.58	-	12.33	-
MDM [92]	-	4.30	45.32	-	6.80	-
DAN [49]	-	3.96	47.00	-	2.67	-
Chen et al. [17]	Hourglass	-	53.64	-	2.50	-
Deng et al. [21]	-	-	-	47.52	-	5.50
Fan et al. [26]	-	-	-	48.02	-	14.83
Dreg + MDM [35]	ResNet101	-	-	52.19	-	3.67
JMFA [22]	Hourglass	-	-	54.85	-	1.00
Our approach	HRNetV2-W18	3.85	52.09	61.55	1.00	0.33

Semantic segmentation

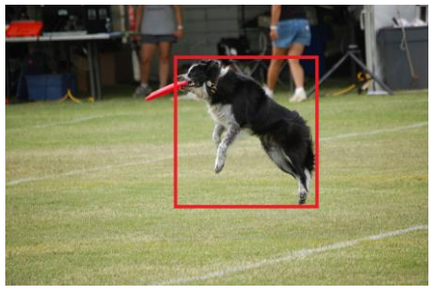
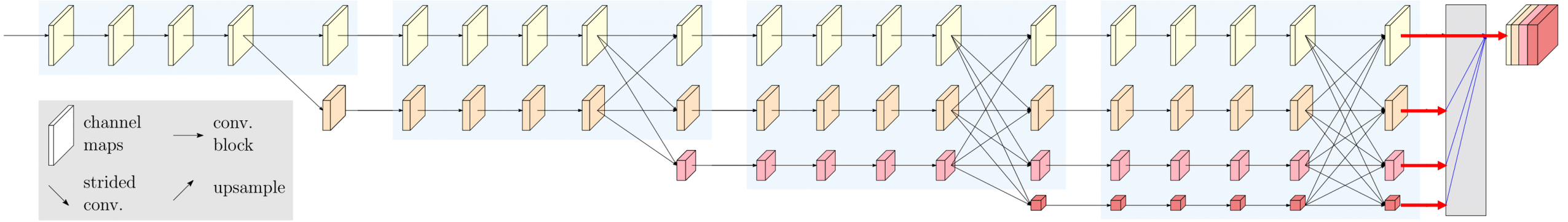




Image classification | Object detection | **Semantic segmentation** | Face alignment | Pose estimation

HRNet for segmentation



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	**82.5**

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



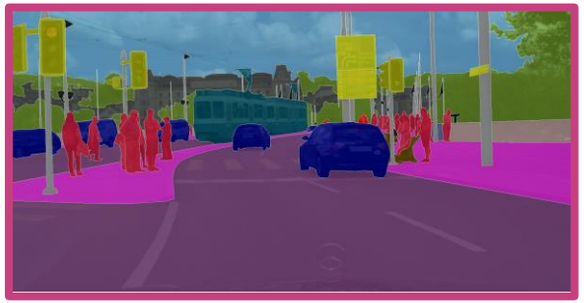
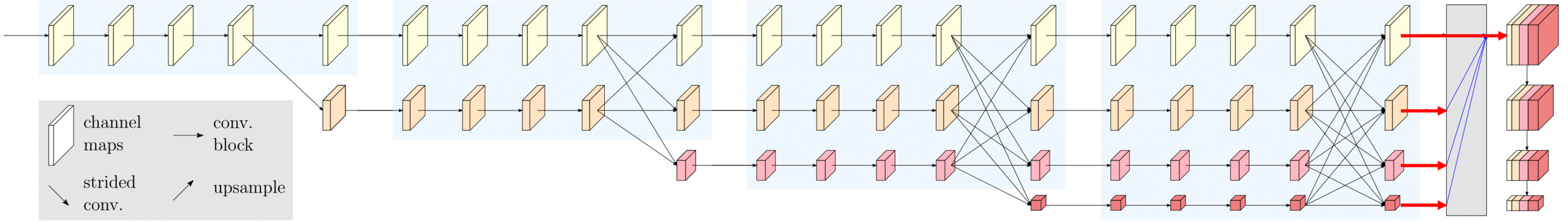


Image classification | **Object detection** | Semantic segmentation | Face alignment | Pose estimation

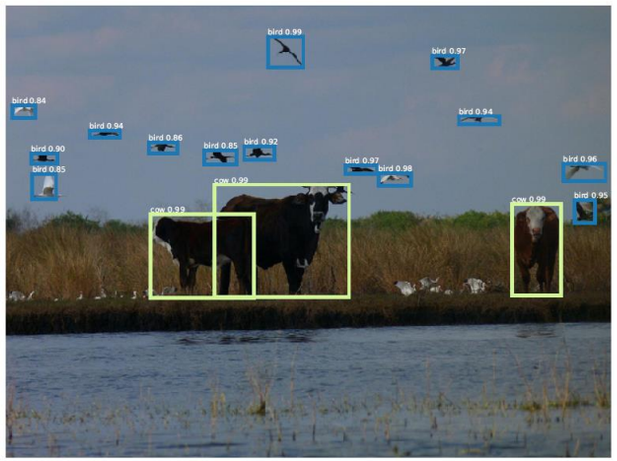
HRNet-FPN for object detection

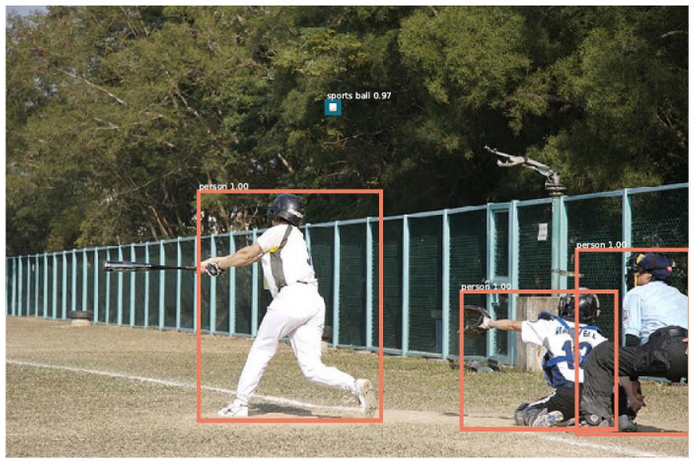


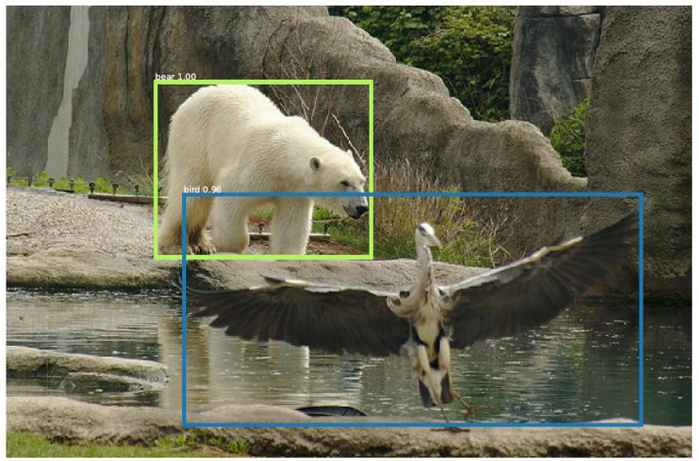
Faster R-CNN

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

single model single scale



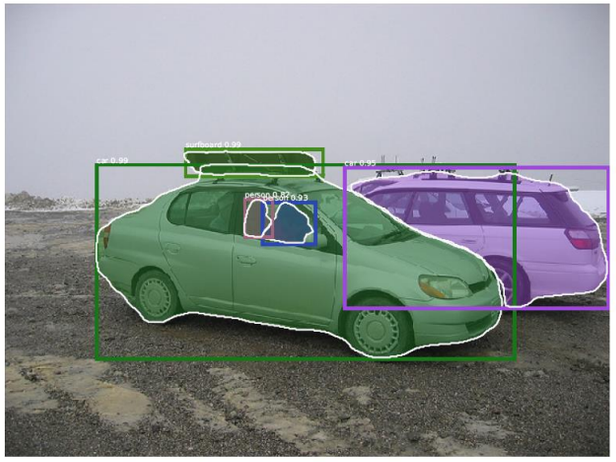
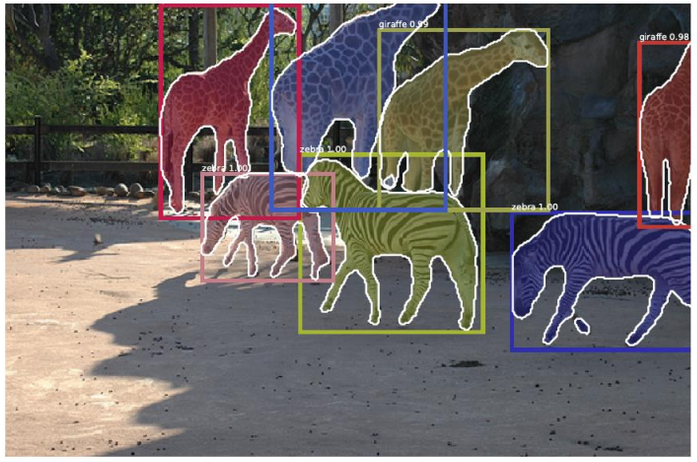




Mask R-CNN

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



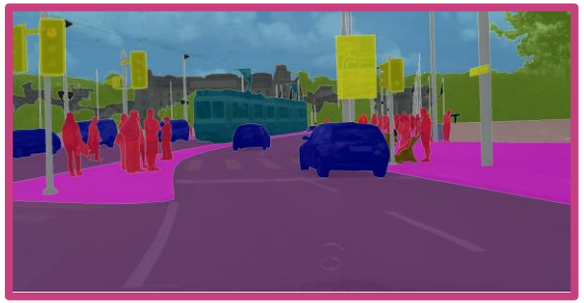


Image classification

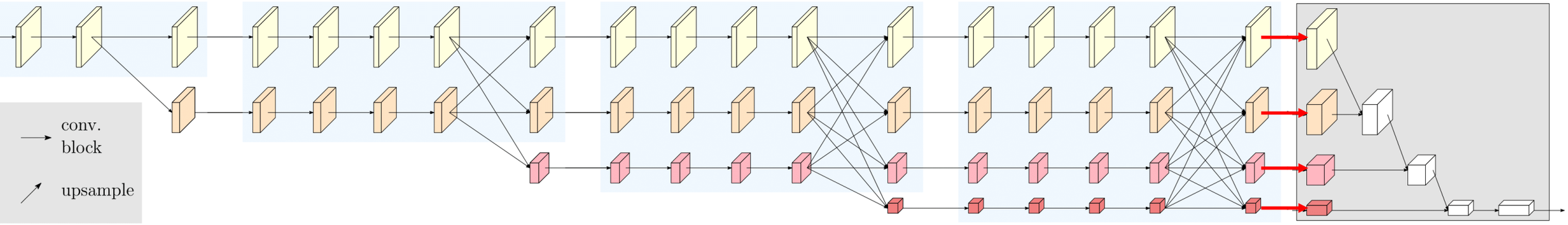
Object detection

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



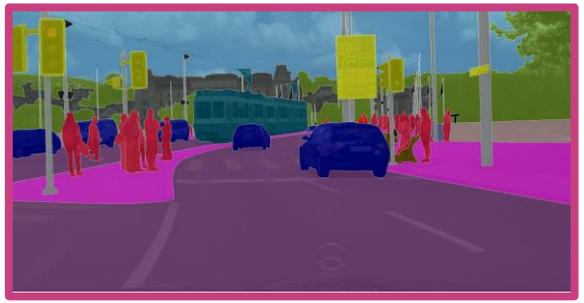


Image classification

Object detection

Semantic segmentation

Face alignment

Pose estimation

Runtime memory and computation cost

	Human pose estimation		Semantic segmentation		
	SB-ResNet-152	HRNet-W48	PSPNet	DeepLabV3	HRNet-W48
train mem	14.8G	7.3.G	14.4G	13.3G	13.9G
infer mem/img	0.29G	0.27G	1.60G	1.15G	1.79G
train sec/iter	1.085	1.231	0.837	0.850	0.692
infer sec/img	0.030 (0.012)	0.058 (0.017)	0.397	0.411	0.150
score	72.0	75.1	79.7	78.5	81.1

Runtime memory and computation cost

	Object detection (Faster R-CNN)			
	ResNet-101	ResNext-101	HRNet-W32	HRNet-W48
train mem	5.4G	9.5G	8.5G	11.3G
infer mem/img	0.62G	0.77G	0.51G	0.79G
train sec/iter	0.550	1.183	0.690	0.965
infer sec/img	0.087	0.144	0.101	0.116
score	39.8	40.8	40.9	41.8

Discussions

high-resolution networks

vs

classification networks

neural architecture design (NAD)

vs

neural architecture search (NAS)

HRNet vs classification network

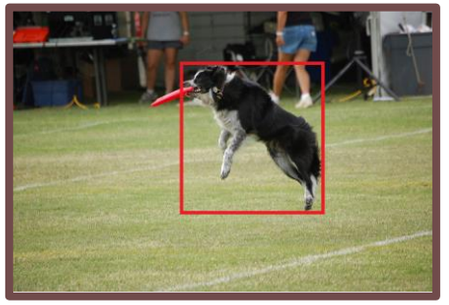




image-level | region-level | pixel-level

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)

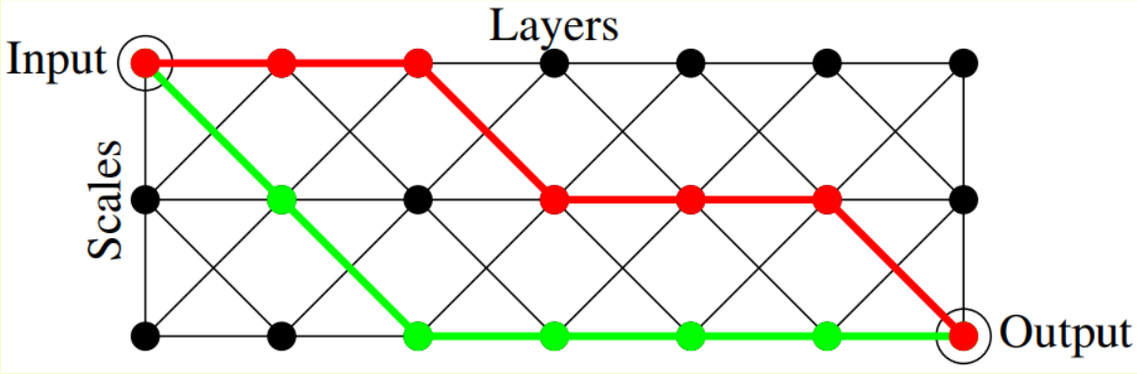
vs

neural architecture search (NAS)

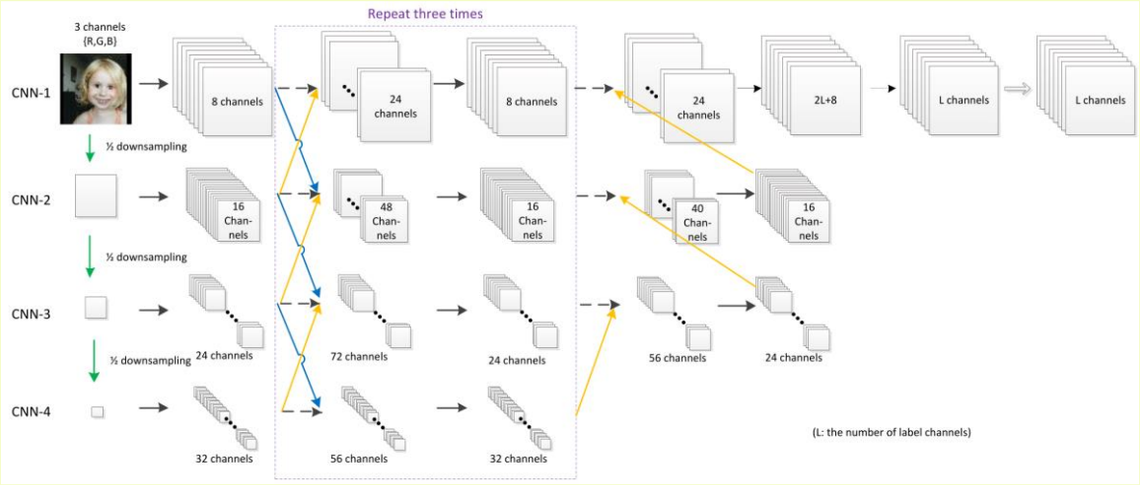
NAD expands search space for NAS

Related networks

Improper fusion frequency

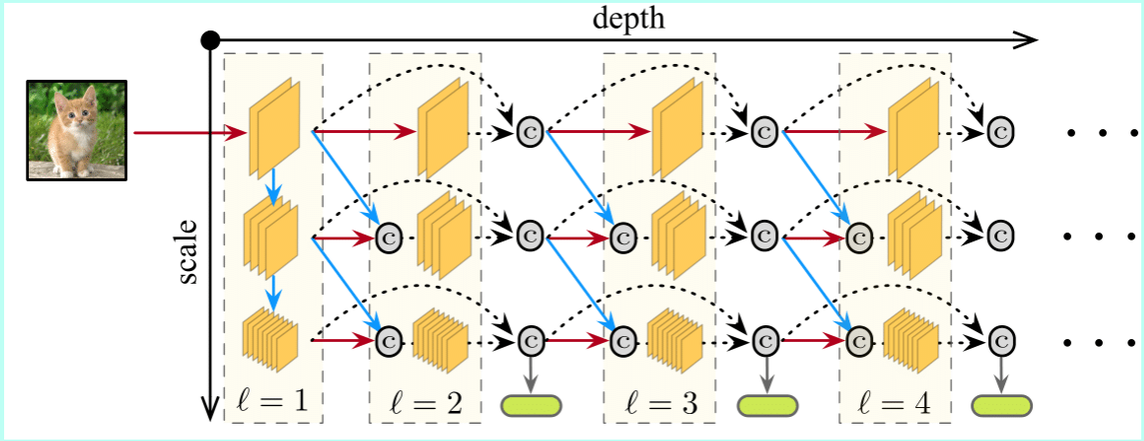


Convolutional neural fabrics

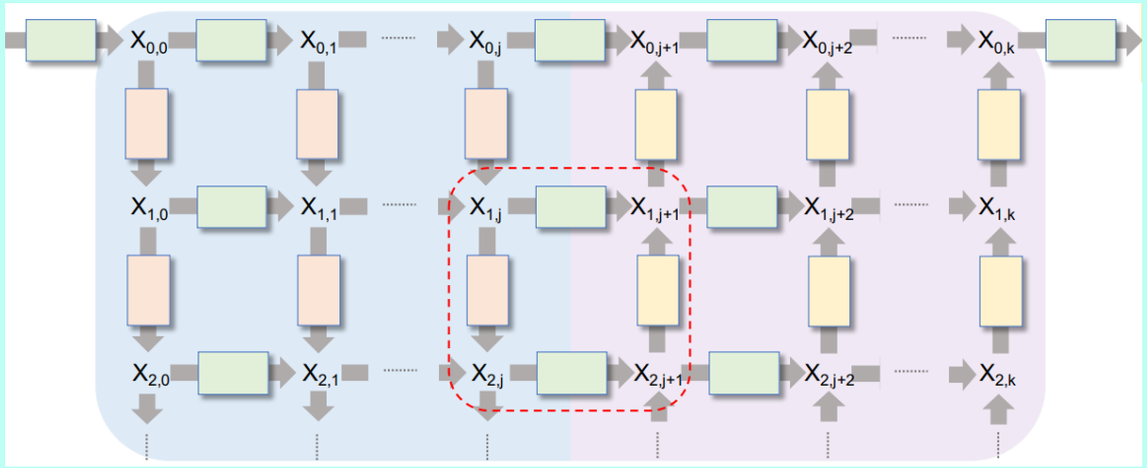


Interlinked CNN

unidirectional fusion



Multi-scale densenet



Gridnet: generalized U-Net

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 OpenImage (ICCV 2019) instance segmentation winner adopted the HRNet
- 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