


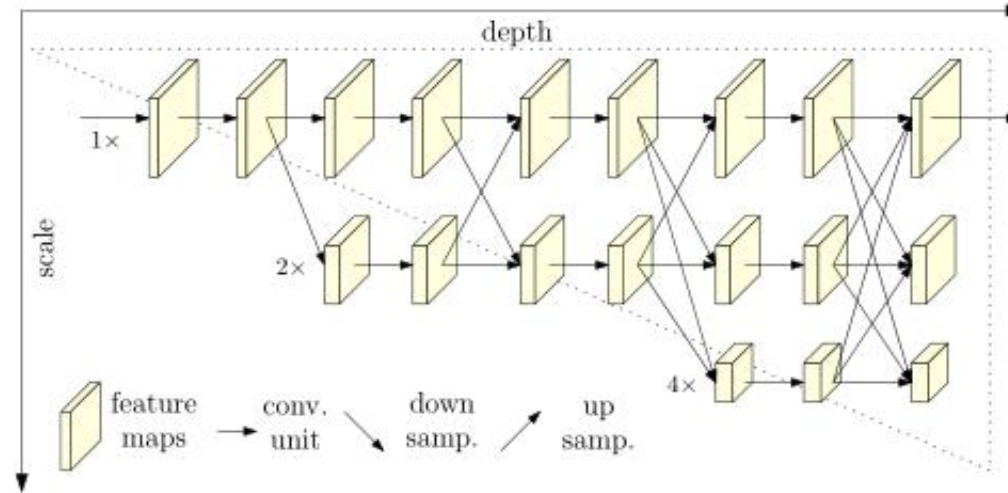
Deep High-Resolution Representation Learning for Human Pose Estimation

yhao.chen0617@gmail.com

1 June, 2019



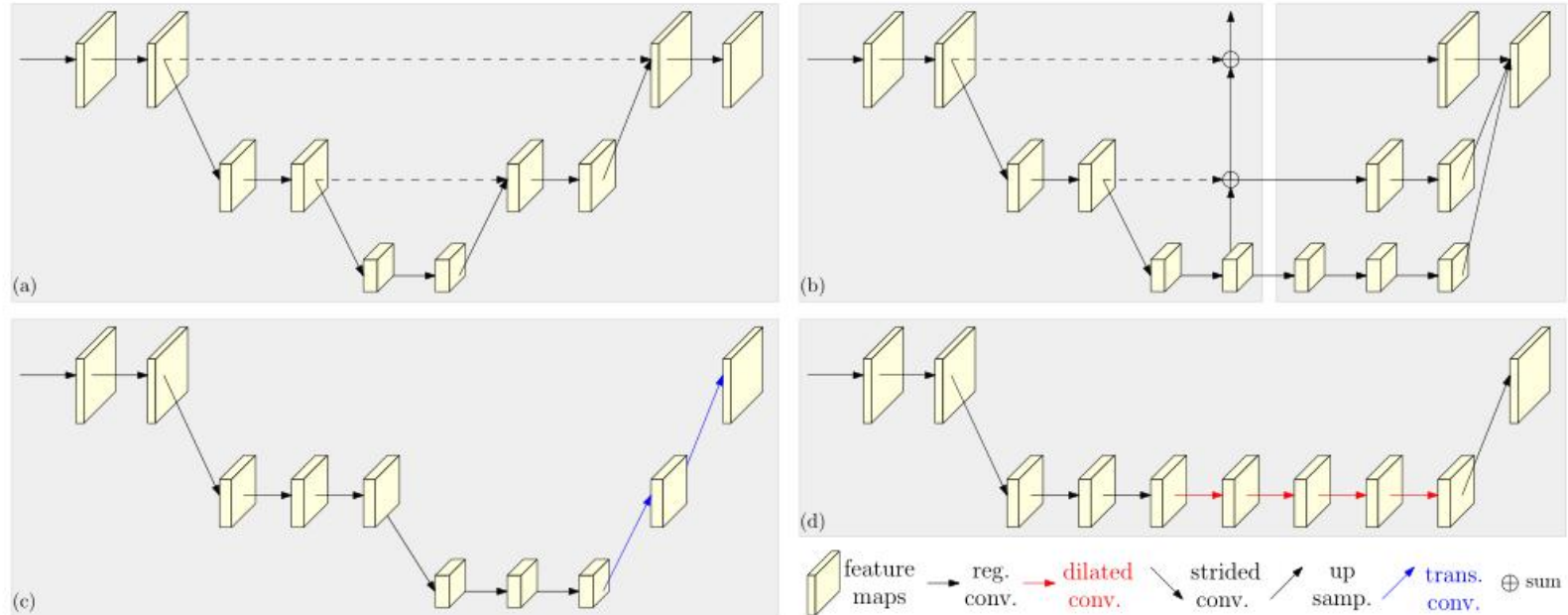
Motivation



In this paper, we are interested in the **human pose estimation** problem with a focus on learning reliable high-resolution representations. Most existing methods **recover high-resolution representations from low-resolution representations** produced by a high-to-low resolution network.

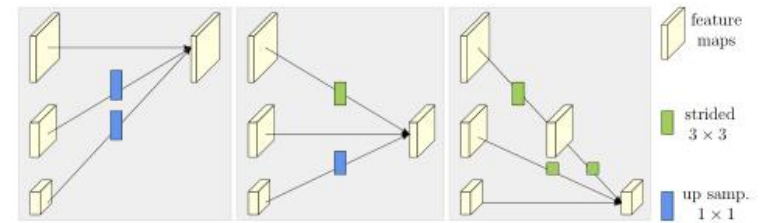
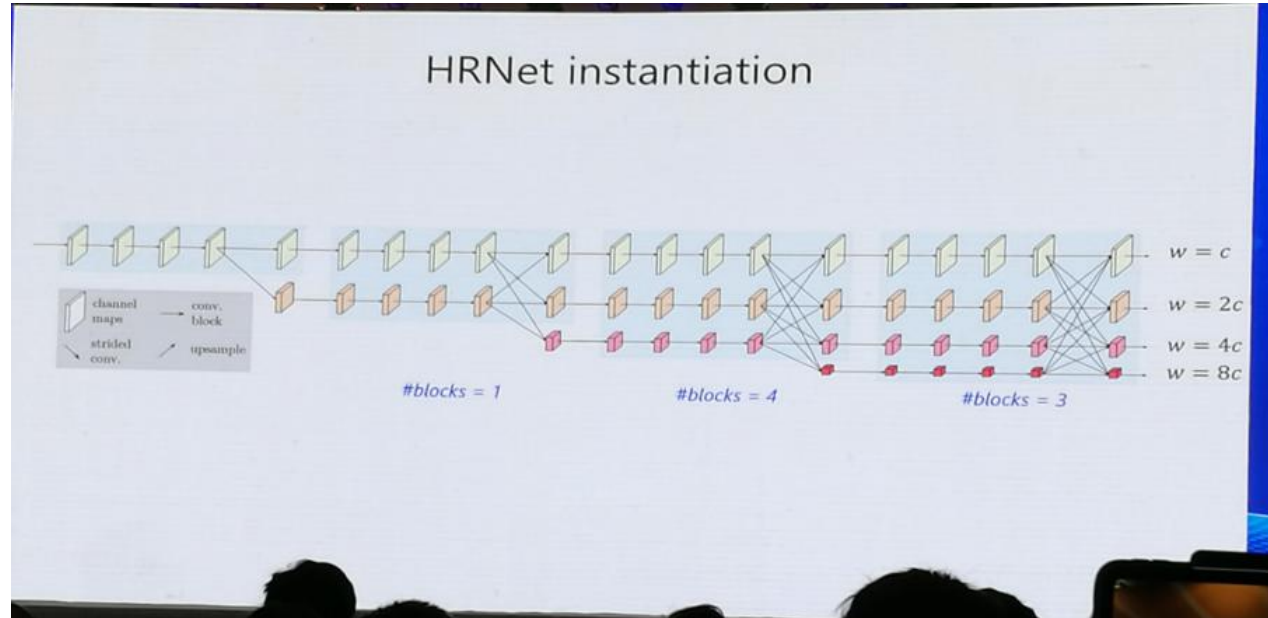
Instead, our proposed network **maintains high-resolution representations** through the whole process.

Related Work



- Connect high-to-low resolution sub-networks in **parallel** rather than in **series**.
- **Maintain** the high resolution instead of **recovering** the resolution through a low-to-high process.
- Most existing fusion schemes **aggregate** low-level and high-level representations. Instead, we perform **repeated** multi-scale fusions.

Approach



- In the exchange unit, each output is an aggregation of the input maps.
- Each stage has multi-times exchange.
- We regress the heatmaps simply from the high-resolution representations output by the last exchange unit

Experiment

Table 1. Comparisons on the COCO validation set. Pretrain = pretrain the backbone on the ImageNet classification task. OHKM = online hard keypoints mining [11].

Method	Backbone	Pretrain	Input size	#Params	GFLOPs	AP	AP ⁵⁰	AP ⁷⁵	AP ^M	AP ^L	AR
8-stage Hourglass [40]	8-stage Hourglass	N	256×192	25.1M	14.3	66.9	—	—	—	—	—
CPN [11]	ResNet-50	Y	256×192	27.0M	6.20	68.6	—	—	—	—	—
CPN + OHKM [11]	ResNet-50	Y	256×192	27.0M	6.20	69.4	—	—	—	—	—
SimpleBaseline [72]	ResNet-50	Y	256×192	34.0M	8.90	70.4	88.6	78.3	67.1	77.2	76.3
SimpleBaseline [72]	ResNet-101	Y	256×192	53.0M	12.4	71.4	89.3	79.3	68.1	78.1	77.1
SimpleBaseline [72]	ResNet-152	Y	256×192	68.6M	15.7	72.0	89.3	79.8	68.7	78.9	77.8
HRNet-W32	HRNet-W32	N	256×192	28.5M	7.10	73.4	89.5	80.7	70.2	80.1	78.9
HRNet-W32	HRNet-W32	Y	256×192	28.5M	7.10	74.4	90.5	81.9	70.8	81.0	79.8
HRNet-W48	HRNet-W48	Y	256×192	63.6M	14.6	75.1	90.6	82.2	71.5	81.8	80.4
SimpleBaseline [72]	ResNet-152	Y	384×288	68.6M	35.6	74.3	89.6	81.1	70.5	79.7	79.7
HRNet-W32	HRNet-W32	Y	384×288	28.5M	16.0	75.8	90.6	82.7	71.9	82.8	81.0
HRNet-W48	HRNet-W48	Y	384×288	63.6M	32.9	76.3	90.8	82.9	72.3	83.4	81.2

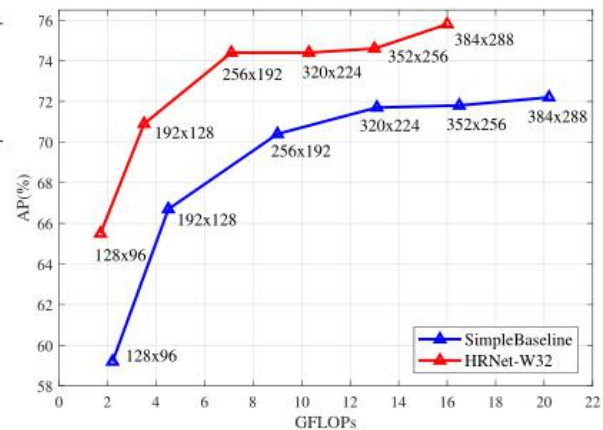


Figure 6. Illustrating how the performances of our HRNet and SimpleBaseline [72] are affected by the input size.

Note

2. Deep High-Resolution Representation Learning for Human Pose Estimation

王井东老师也是在workshop中力推了这篇文章，而且一做Ke Sun在poster环节也展示了这篇论文。

论文的motivation还是很easy的，我们就是想要获得高分辨率高语义的feature。现有的大部分方法，基于hourglass都是从low-resolution的feature中恢复到high-resolution，在恢复过程中难免会有一些问题，这也是必然的，那很显然的就是说能不能一直保留high resolution的feature，贯穿整个cnn model。

盗用王井东老师的一张PPT，网络的结构大致如上图所示，在每个stage结束后，每个分辨率的feature都是由现有的所有的分辨率的feature共同获得的，上采样用最近邻和 1×1 conv，下采样用 3×3 s=2，ke sun说上下采样他们自己试过很多方法，包括pooling，bilinear等，发现这样设置是最好用的。当用于pose时，合并最后一个stage的所有feature得到高分辨的feature，用于最后的检测。做detection时，用合并到的最后的高-resolution的feature下采样产生一组不同scale的feature，构成feature pyramid。网络整个flops不会太大，因为高分辨率的feature channel很少，合并不同分辨率的feature用的sum，不是concat，不改变channel数目。其实个人估计网络的速度也不会太快，因为高分辨率的feature太多了，会导致mac很高，从而降低速度（参考shufflenet v2），而且网络中的连接太多。

conclusion: 1.其实HRNet的motivation真的不要太简单，但是在性能上也不要太work，很棒的工作~ 从结构上看HRNet更像是Unet的极限连接版。这么来说的话，是不是去反思一下以前的经典结构，来个rethink，可以发一篇顶会？哈哈哈；

2. 在应用到detection的时候，其实构造的fp各层的knowledge是不一样的，虽然现在的大尺度的feature也是high-level的，但是这没有触及到detection中的scale问题的本质，这是一个接下来想考虑的方向；

3. 类似于densenet的后续改进，hrnet中的这么多连接真的有必要吗？如何有效的去除冗余性。

How to expand HRNet to labeling image/region/pixel tasks ?

HRNet V2

Approach

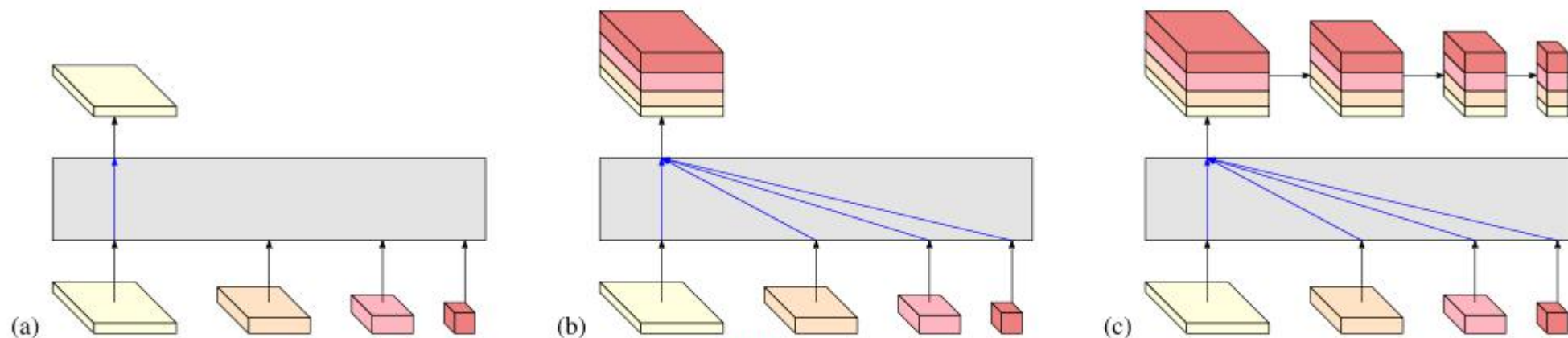
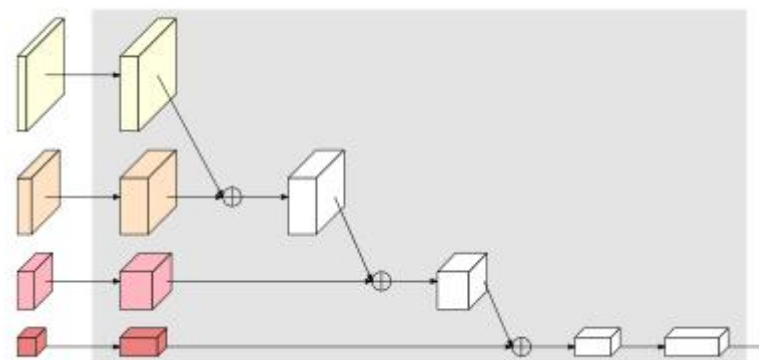


Figure 3. (a) The high-resolution representation proposed in [91] (HRNetV1); (b) Concatenating the (upsampled) representations that are from all the resolutions for semantic segmentation and facial landmark detection (HRNetV2); (c) A feature pyramid formed over (b) for object detection (HRNetV2p). The four-resolution representations at the bottom in each sub-figure are outputted from the network in Figure 1, and the gray box indicates how the output representation is obtained from the input four-resolution representations.

Aggregating the (upsampled) representations from all the parallel convolutions rather than **only** the representation from the high-resolution convolution.



Experiment

Table 1. Segmentation results on Cityscapes val (single scale and no flipping). The GFLOPs is calculated on the input size 1024×2048 .

	backbone	#param.	GFLOPs	mIoU
UNet++ [133]	ResNet-101	59.5M	748.5	75.5
DeepLabv3 [14]	Dilated-ResNet-101	58.0M	1778.7	78.5
DeepLabv3+ [16]	Dilated-Xception-71	43.5M	1444.6	79.6
PSPNet [126]	Dilated-ResNet-101	65.9M	2017.6	79.7
Our approach	HRNetV2-W40	45.2M	493.2	80.2
Our approach	HRNetV2-W48	65.9M	747.3	81.1

Table 6. Object detection results evaluated on COCO val in the Faster R-CNN framework. LS = learning schedule.

backbone	LS	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
ResNet-50-FPN	1×	36.7	58.3	39.9	20.9	39.8	47.9
HRNetV2p-W18	1×	36.2	57.3	39.3	20.7	39.0	46.8
ResNet-50-FPN	2×	37.6	58.7	41.3	21.4	40.8	49.7
HRNetV2p-W18	2×	38.0	58.9	41.5	22.6	40.8	49.6
ResNet-101-FPN	1×	39.2	61.1	43.0	22.3	42.9	50.9
HRNetV2p-W32	1×	39.6	61.0	43.3	23.7	42.5	50.5
ResNet-101-FPN	2×	39.8	61.4	43.4	22.9	43.6	52.4
HRNetV2p-W32	2×	40.9	61.8	44.8	24.4	43.7	53.3
ResNet-152-FPN	1×	39.5	61.2	43.0	22.1	43.3	51.8
HRNetV2p-W40	1×	40.4	61.8	44.1	23.8	43.8	52.3
ResNet-152-FPN	2×	40.6	61.9	44.5	22.8	44.0	53.1
HRNetV2p-W40	2×	41.6	62.5	45.6	23.8	44.9	53.8
X-101-64×4d-FPN	1×	41.3	63.4	45.2	24.5	45.8	53.3
HRNetV2p-W48	1×	41.3	62.8	45.1	25.1	44.5	52.9
X-101-64×4d-FPN	2×	40.8	62.1	44.6	23.2	44.5	53.7
HRNetV2p-W48	2×	41.8	62.8	45.9	25.0	44.7	54.6

Table 14. ImageNet Classification results of HRNet and ResNets. The proposed method is named HRNet-W x -C.

	#Params.	GFLOPs	top-1 err.	top-5 err.
<i>Residual branch formed by two 3×3 convolutions</i>				
ResNet-38	28.3M	3.80	24.6%	7.4%
HRNet-W18-C	21.3M	3.99	23.1%	6.5%
ResNet-72	48.4M	7.46	23.3%	6.7%
HRNet-W30-C	37.7M	7.55	21.9%	5.9%
ResNet-106	64.9M	11.1	22.7%	6.4%
HRNet-W40-C	57.6M	11.8	21.1%	5.6%
<i>Residual branch formed by a bottleneck</i>				
ResNet-50	25.6M	3.82	23.3%	6.6%
HRNet-W44-C	21.9M	3.90	23.0%	6.5%
ResNet-101	44.6M	7.30	21.6%	5.8%
HRNet-W76-C	40.8M	7.30	21.5%	5.8%
ResNet-152	60.2M	10.7	21.2%	5.7%
HRNet-W96-C	57.5M	10.2	21.0%	5.6%

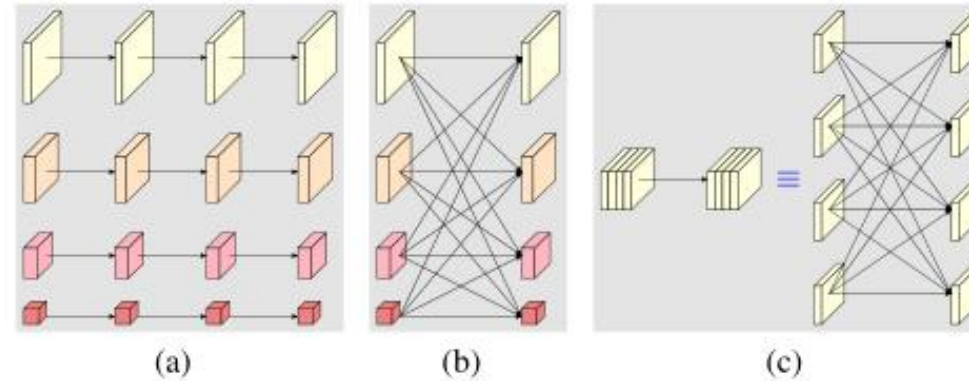


Figure 2. Multi-resolution block: (a) multi-resolution group convolution and (b) multi-resolution convolution. (c) A normal convolution (left) is equivalent to fully-connected multi-branch convolutions (right).

The 2nd, 3rd and 4th stages are formed by repeating modularized **multi-resolution blocks**. A multi-resolution block consists of a multi-resolution group convolution and a multi-resolution convolution.

Thank you for attention to everyone!

2 June, 2019