Deep High-Resolution Representation Learning for Visual Recognition

姓名：魏子继

2023年11月25日

中国科学院大学
University of Chinese Academy of Sciences

文章简介

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, MARCH 2020 1

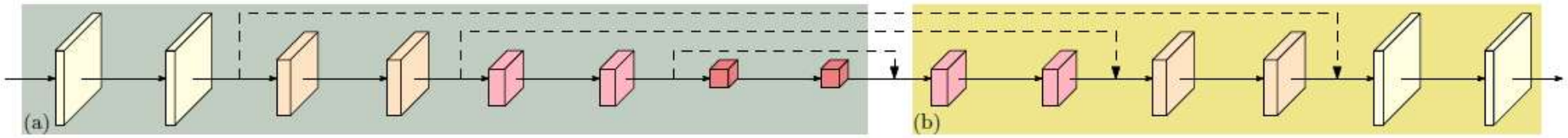
Deep High-Resolution Representation Learning for Visual Recognition

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

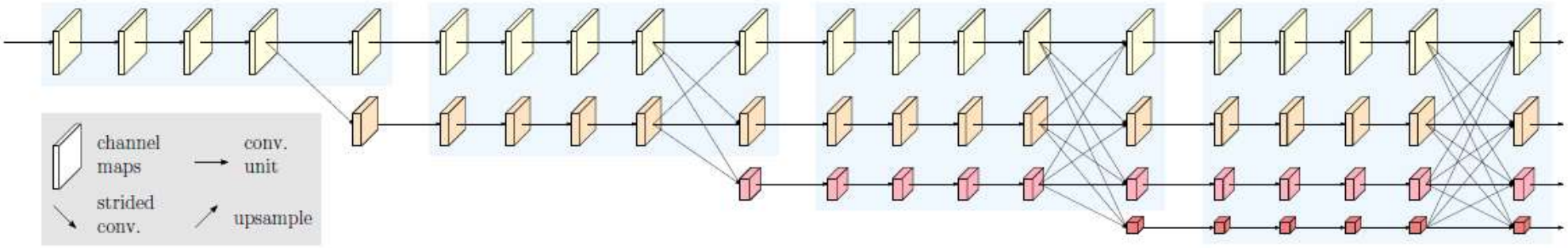
研究意义：处理过程中维持图像的高分辨率表达，使结果拥有更加丰富的情景信息和更加精确的空间表达

Cite: J. Wang *et al*., "Deep High-Resolution Representation Learning for Visual Recognition," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3349-3364, 1 Oct. 2021, doi: 10.1109/TPAMI.2020.2983686.

框架对比



常用的encoder-decoder网络框架：从低分辨率图像中恢复高分辨率图像

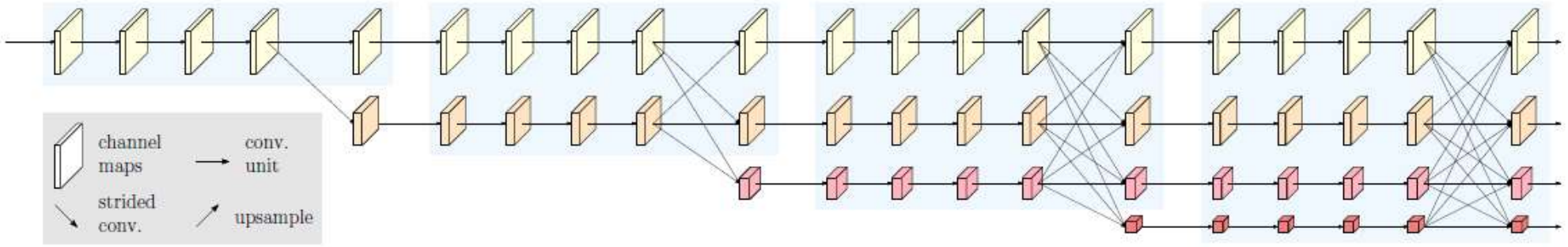


本文提出的HRNet网络框架：并行连接、不断融合、始终保持高分辨率表达

doi: 10.1109/TPAMI.2020.2983686.

实施细节

并行多分辨率卷积：



$$\begin{array}{ccccc} \mathcal{N}_{11} & \rightarrow & \mathcal{N}_{21} & \rightarrow & \mathcal{N}_{31} & \rightarrow & \mathcal{N}_{41} \\ & \searrow & \mathcal{N}_{22} & \rightarrow & \mathcal{N}_{32} & \rightarrow & \mathcal{N}_{42} \\ & & & \searrow & \mathcal{N}_{33} & \rightarrow & \mathcal{N}_{43} \\ & & & & & \searrow & \mathcal{N}_{44}, \end{array}$$

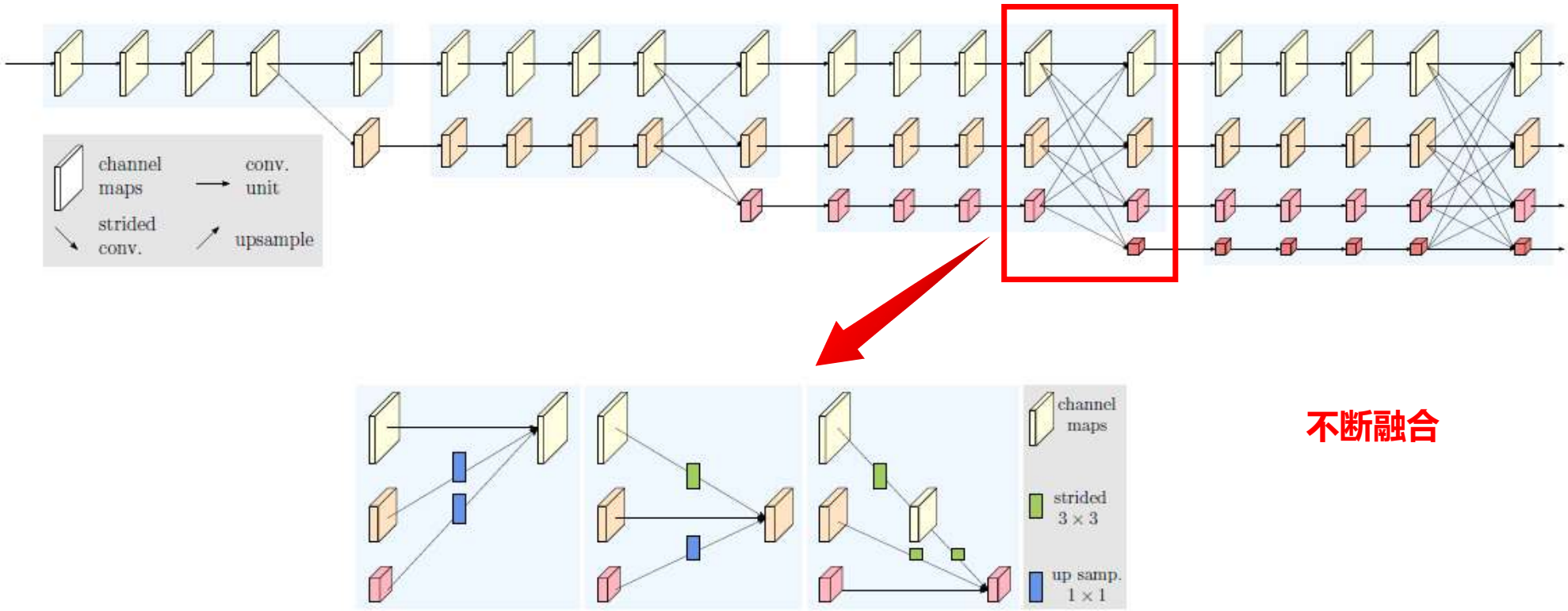
并行连接

始终保持高分辨率表达

doi: 10.1109/TPAMI.2020.2983686.

实施细节

重复多分辨率融合：



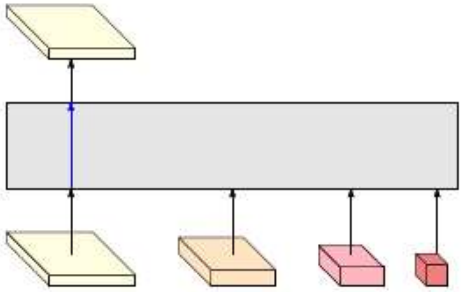
不断融合

doi: 10.1109/TPAMI.2020.2983686.

实施细节

HRNet表达头：

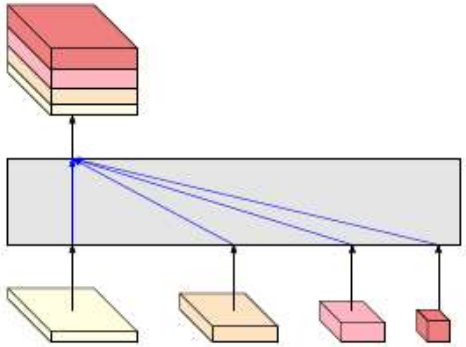
人体姿态估计



HRVNetV1

只输出高分辨率卷积流计算得到的高分辨率特征

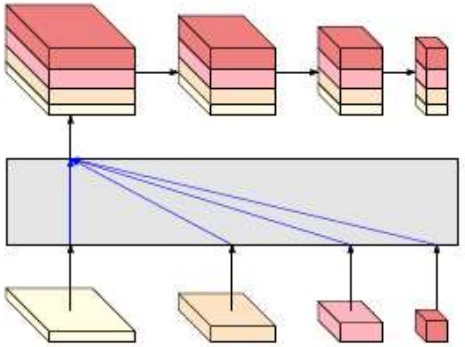
图像分割



HRVNetV2

串联所有高分辨率到低分辨率并行流的输出特征

目标检测



HRVNetV2p

降采样V2输出结果，构建特征金字塔**多层次表达**

doi: 10.1109/TPAMI.2020.2983686.

实施细节

HRNet框架实施：

Resolution	Stage 1	Stage 2	Stage 3	Stage 4
4×	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 4 \times 1$	$\begin{bmatrix} 3\times3, C \\ 3\times3, C \end{bmatrix} \times 4 \times 1$	$\begin{bmatrix} 3\times3, C \\ 3\times3, C \end{bmatrix} \times 4 \times 4$	$\begin{bmatrix} 3\times3, C \\ 3\times3, C \end{bmatrix} \times 4 \times 3$
8×		$\begin{bmatrix} 3\times3, 2C \\ 3\times3, 2C \end{bmatrix} \times 4 \times 1$	$\begin{bmatrix} 3\times3, 2C \\ 3\times3, 2C \end{bmatrix} \times 4 \times 4$	$\begin{bmatrix} 3\times3, 2C \\ 3\times3, 2C \end{bmatrix} \times 4 \times 3$
16×			$\begin{bmatrix} 3\times3, 4C \\ 3\times3, 4C \end{bmatrix} \times 4 \times 4$	$\begin{bmatrix} 3\times3, 4C \\ 3\times3, 4C \end{bmatrix} \times 4 \times 3$
32×				$\begin{bmatrix} 3\times3, 8C \\ 3\times3, 8C \end{bmatrix} \times 4 \times 3$

doi: 10.1109/TPAMI.2020.2983686.

实验结果

人体关键点检测：

TABLE 2
Comparisons on COCO `test-dev`. The observations are similar to the results on COCO `val`.

Method	Backbone	Input size	#Params	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
Bottom-up: keypoint detection and grouping										
OpenPose [15]	–	–	–	–	61.8	84.9	67.5	57.1	68.2	66.5
Associative Embedding [104]	–	–	–	–	65.5	86.8	72.3	60.6	72.6	70.2
PersonLab [108]	–	–	–	–	68.7	89.0	75.4	64.1	75.5	75.4
MultiPoseNet [72]	–	–	–	–	69.6	86.3	76.6	65.0	76.3	73.5
Top-down: human detection and single-person keypoint detection										
Mask-RCNN [53]	ResNet-50-FPN	–	–	–	63.1	87.3	68.7	57.8	71.4	–
G-RMI [109]	ResNet-101	353 × 257	42.6M	57.0	64.9	85.5	71.3	62.3	70.0	69.7
Integral Pose Regression [132]	ResNet-101	256 × 256	45.0M	11.0	67.8	88.2	74.8	63.9	74.0	–
G-RMI + extra data [109]	ResNet-101	353 × 257	42.6M	57.0	68.5	87.1	75.5	65.8	73.3	73.3
CPN [24]	ResNet-Inception	384 × 288	–	–	72.1	91.4	80.0	68.7	77.2	78.5
RMPE [38]	PyraNet [165]	320 × 256	28.1M	26.7	72.3	89.2	79.1	68.0	78.6	–
CFN [60]	–	–	–	–	72.6	86.1	69.7	78.3	64.1	–
CPN (ensemble) [24]	ResNet-Inception	384 × 288	–	–	73.0	91.7	80.9	69.5	78.1	79.0
SimpleBaseline [152]	ResNet-152	384 × 288	68.6M	35.6	73.7	91.9	81.1	70.3	80.0	79.0
HRNetV1	HRNetV1-W32	384 × 288	28.5M	16.0	74.9	92.5	82.8	71.3	80.9	80.1
HRNetV1	HRNetV1-W48	384 × 288	63.6M	32.9	75.5	92.5	83.3	71.9	81.5	80.5
HRNetV1 + extra data	HRNetV1-W48	384 × 288	63.6M	32.9	**77.0**	**92.7**	**84.5**	**73.4**	**83.1**	**82.0**

doi: 10.1109/TPAMI.2020.2983686.

实验结果

人体关键点检测：

Fig. 6. Qualitative COCO human pose estimation results over representative images with various human size, different poses, or clutter background.

明显优于自底向上的方法

小网络在检测精度、模型尺寸和计算复杂度方面更有优势

大网络在相同设置下检测精度更高

doi: 10.1109/TPAMI.2020.2983686.

实验结果

图像分割：

Fig. 7. Qualitative segmentation examples from Cityscapes (left two), PASCAL-Context (middle two), and LIP (right two).

TABLE 6
Semantic segmentation results on LIP. Our method doesn't exploit any extra information, e.g., pose or edge. The overall performance of our approach is the best, and the OCR scheme [170] further improves the segmentation quality. D-ResNet-101 = Dilated-ResNet-101.

	backbone	extra.	pixel acc.	avg. acc.	mIoU
Attention+SSL [47]	VGG16	Pose	84.36	54.94	44.73
DeepLabV3+ [22]	D-ResNet-101	-	84.09	55.62	44.80
MMAN [100]	D-ResNet-101	-	-	-	46.81
SS-NAN [183]	ResNet-101	Pose	87.59	56.03	47.92
MuLA [106]	Hourglass	Pose	**88.50**	60.50	49.30
JPPNet [87]	D-ResNet-101	Pose	86.39	62.32	51.37
CE2P [98]	D-ResNet-101	Edge	87.37	63.20	53.10
HRNetV2	HRNetV2-W48	N	88.21	67.43	55.90
HRNetV2 + OCR [170]	HRNetV2-W48	N	88.24	**67.84**	**56.48**

Cityscapes：更小的模型尺寸、更低的计算复杂度、更高的检测精度

PASCAL-Context：配合OCR主题检测精度更高

LIP：不需要使用任何额外的信息、更小的模型尺寸、更低的计算复杂度、更高的检测精度

doi: 10.1109/TPAMI.2020.2983686.

实验结果

目标检测：

TABLE 7
GFLOPs and #parameters for COCO object detection. The numbers are obtained with the input size 800×1200 and if applicable 512 proposals fed into R-CNN except the numbers for CenterNet are obtained with the input size 511×511. R-x = ResNet-x-FPN, X-101 = ResNeXt-101-64×4d, H-x = HRNetV2p-Wx, and HG-52 = Hourglass-52.

	Faster R-CNN [53]						Cascade R-CNN [13]						FCOS [136]				CenterNet [36]			
	R-50	H-18	R-101	H-32	X-101	H-48	R-50	H-18	R-101	H-32	X-101	H-48	R-50	H-18	R-101	H-32	HG-52	H-48	HG-104	H-64
#param. (M)	39.8	26.2	57.8	45.0	94.9	79.4	69.4	55.1	88.4	74.9	127.3	111.0	32.0	17.5	51.0	37.3	104.8	73.6	210.1	127.7
GFLOPs	172.3	159.1	239.4	245.3	381.8	399.1	226.2	207.8	298.7	300.8	448.3	466.5	190.0	180.3	261.2	273.3	227.0	217.1	388.4	318.5
	Cascade Mask R-CNN [13]						Hybrid Task Cascade [16]						Mask R-CNN [53]							
	R-50	H-18	R-101	H-32	X-101	H-48	R-50	H-18	R-101	H-32	X-101	H-48	R-50	H-18	R-101	H-32				
#param. (M)	77.3	63.1	96.3	82.9	135.2	118.9	80.3	66.1	99.3	85.9	138.2	121.9	44.4	30.1	63.4	49.9				
GFLOPs	431.7	413.1	504.1	506.2	653.7	671.9	476.9	458.3	549.2	551.4	698.9	717.0	266.5	247.9	338.8	341.0				

Cascade R-CNN [12]*	ResNet-101-FPN	800	~ 1.6×	42.8	62.1	46.3	23.7	45.5	55.2
Cascade R-CNN	ResNet-101-FPN	800	~ 1.6×	43.1	61.7	46.7	24.1	45.9	55.0
Cascade R-CNN	HRNetV2p-W32	800	~ 1.6×	43.7	62.0	47.4	25.5	46.0	55.3
Cascade R-CNN	X-101-64 × 4d-FPN	800	~ 1.6×	44.9	63.7	48.9	25.9	47.7	57.1
Cascade R-CNN	HRNetV2p-W48	800	~ 1.6×	44.8	63.1	48.6	26.0	47.3	56.3

注：此为部分检测结果，但具有代表性

doi: 10.1109/TPAMI.2020.2983686.

实验结果

目标检测：

Fig. 8. Qualitative examples for COCO object detection (left three) and instance segmentation (right three).

在相似的模型尺寸和计算复杂度下，HRNet优于ResNet

没有多尺度训练与检测下，在大多数主流框架中， HRNet优于ResNet

doi: 10.1109/TPAMI.2020.2983686.

消融实验

不同分辨率表达：

高分辨率

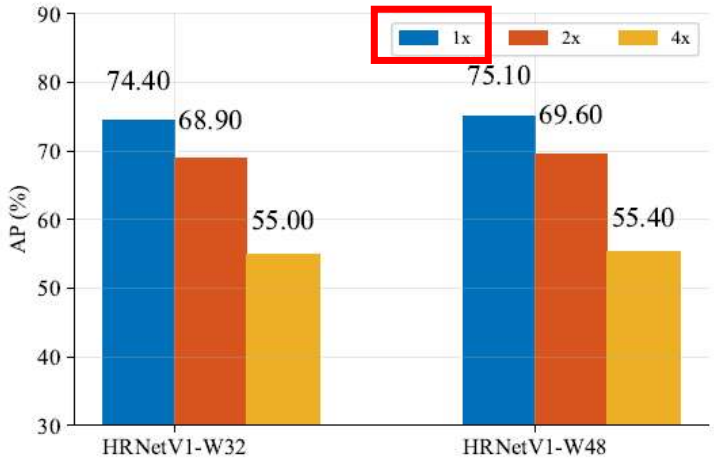


Fig. 9. Ablation study about the resolutions of the representations for human pose estimation. $1\times$, $2\times$, $4\times$ correspond to the representations of the high, medium, low resolutions, respectively. The results imply that higher resolution improves the performance.

分辨率影响关键点检测的质量

重复的多分辨率融合：

TABLE 12
Ablation study for multi-resolution fusion units on COCO `val` human pose estimation (AP) and Cityscapes `val` semantic segmentation (mIoU). Final = final fusion immediately before representation head, Across = intermediate fusions across stages, Within = intermediate fusions within stages. We can see that the three fusions are beneficial for both human pose estimation and semantic segmentation.

Method	Final	Across	Within	Pose (AP)	Segmentation (mIoU)
(a)	✓			70.8	74.8
(b)	✓	✓		71.9	75.4
(c)	✓	✓	✓	73.4	76.4

多分辨率融合单元是有益的

更多的融合将得到更优的效果

doi: 10.1109/TPAMI.2020.2983686.

消融实验

分辨率维护：

Resolution maintenance. We study the performance of a variant of the HRNet: all the four high-to-low resolution streams are added at the beginning and the depths of the four streams are the same; the fusion schemes are the same to ours. Both the HRNets and the variants (with similar #Params and GFLOPs) are trained from scratch.

The human pose estimation performance (AP) on COCO `val` for the variant is 72.5, which is lower than 73.4 for HRNetV1-W32. The segmentation performance (mIoU) on Cityscapes `val` for the variant is 75.7, which is lower than 76.4 for HRNetV2-W48. We believe that the reason is that the low-level features extracted from the early stages over the low-resolution streams are less helpful. In addition, another simple variant, only the high-resolution stream of similar #parameters and GFLOPs without low-resolution parallel streams shows much lower performance on COCO and Cityscapes.

网络早期低分辨率卷积流得到的

低分辨率特征鲜有帮助

V1 vs. V2：

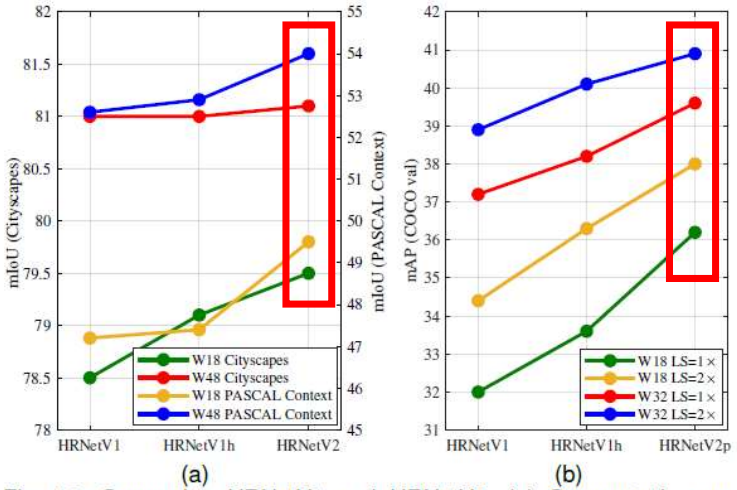


Fig. 10. Comparing HRNetV1 and HRNetV2. (a) Segmentation on Cityscapes `val` and PASCAL-Context for comparing HRNetV1 and its variant HRNetV1h, and HRNetV2 (single scale and no flipping). (b) Object detection on COCO `val` for comparing HRNetV1 and its variant HRNetV1h, and HRNetV2p (LS = learning schedule). We can see that HRNetV2 is superior to HRNetV1 for both semantic segmentation and object detection.

HRNetV2效果明显优于HRNetV1

聚合低分辨率并行卷积流的特征有助于检测性能的提高

doi: 10.1109/TPAMI.2020.2983686.

总结展望

HRNet总结：

通过并行连接高分辨率到低分辨率卷积流，同时不断融合各卷积流中的信息，使处理过程中维持图像的高分辨率表达，达到检测结果拥有更加丰富的情景信息和更加精确的空间表达的目的

HRNet展望：

鼓励更多研究向通过直接设计神经网络框架解决特定视觉问题方面开展

研究在图像分割领域，HRNet与其他技术的融合，如与OCR技术结合

研究在计算机视觉领域，HRNet更加广泛的应用，如无人机目标检测

doi: 10.1109/TPAMI.2020.2983686.

THANKS

姓名：魏子继

2023年11月25日

中国科学院大学

University of Chinese Academy of Sciences