图像处理

文献阅读

姓名：魏子继　　学号：202318019427048

Title: Deep High-Resolution Representation Learning for Visual Recognition

本文提出一种名为 High-Resolution Network(HRNet)的新型视觉识别框架，该框架旨在通过并行连接高分辨率到低分辨率的卷积流，同时在不同分辨率之间重复交换信息，以保持图像在处理全过程中保持高分辨率表示。与传统的编码-解码网络框架相比，HRNet 的最终输出表示包含更丰富的语义信息和更精确的空间定位。

1 网络背景

当前，凭借能够挖掘更加丰富的语义信息，深度卷积神经网络被广泛应用在图像分割任务中。基于 encoder-decoder 模型的方法是当前主流的深度学习图像分割方法之一，先前的基于 encoder-decoder 模型的深度网络结构如图 1-1 所示，该网络通过串行连接的方式，首先将高分辨率图像下采样到低分辨率图像，得到影像的低分辨率特征表达，在此基础上完成分类、分割等任务的处理后，将结果通过上采样的方式逐渐恢复为高分辨率图像。

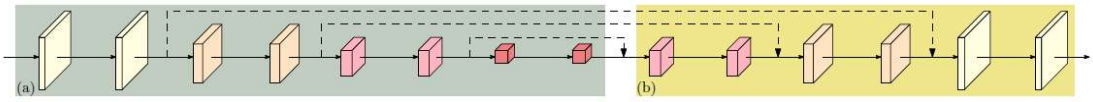

图 1-1：传统的基于 encoder-decoder 模型

该论文提出一种新型的轻量级网络框架 HRNet，输出结果特征包含更加丰富的语义信息和更加精确的空间定位。HRNet 的网络结构如图 1-2 所示，区别于传统的基于 encoder-decoder 模型结构，HRNet 采用并行连接的方式，从高分辨率卷积流出发，逐渐并行地增加低分辨率卷积流，并在每一并行卷积流阶段的末尾进行不同分辨率特征之间的信息融合，以实现在网络的整个处理过程中维持图像的高分辨率表达，从而获得具有较强空间位置敏感性的输出特征。

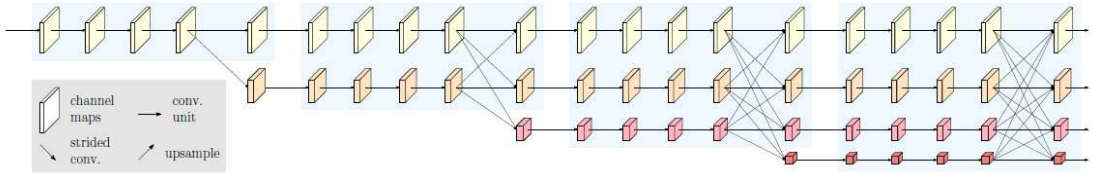

图 1-2：HRNet 网络结构

2 网络实施细节

HRNet 框架结构包含并行多分辨率卷积、重复多分辨率融合、不同表达头的特点。

并行分辨率卷积的一种示例图如图 2-1 所示。该卷积结构从高分辨率卷积流出发，作为第一阶段，记为$\mathcal{N}_{11}$。在其后逐渐增加较低分辨率的卷积流，并通过并行连接的方式将同阶段中不同分辨率的卷积流连接，以形成新的卷积阶段，从而形成整个卷积网络。因此，各阶段中并行连接的卷积流由先前阶段的卷积流和更低一档分辨率的卷积流组成。以第二阶段为例，$\mathcal{N}_{21}$拥有与$\mathcal{N}_{11}$相同的高分辨率，$\mathcal{N}_{22}$的分辨率比$\mathcal{N}_{21}$的分辨率低一档，两个不同分辨率的卷积流通过并行的方式连接，在其后第三阶段的并行卷积中，$\mathcal{N}_{31}$与$\mathcal{N}_{21}$、$\mathcal{N}_{32}$与$\mathcal{N}_{22}$具有相同分辨率，$\mathcal{N}_{33}$的分辨率比$\mathcal{N}_{32}$的分辨率第一档，如此反复并行，最终形成整个并行网络。

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

图 2-1：并行分辨率卷积示意图

重复多分辨率融合的目标是交换多分辨率表达之间的信息，其示例图如图 2-2 所示。在并行卷积结构的每个阶段末尾，将进行多分辨率融合的工作，前三个阶段的融合是三个输入映射的总和，第四个阶段的融合将拥有一个额外的输出。其中，映射函数的选取与卷积流的分辨率有关，当输入与输出的分辨率相等时，映射的结果与输入相等；当输入分辨率小于输出分辨率时，通过双线性插值的上采样方式将输入映射至与输出分辨率相等，例如，上采样通过 1*1 的卷积实现；当输入分辨率大于输出分辨率时，通过下采样的方式将输入映射至与输出分辨率相等，例如，下采样通过 3*3 的 strided 实现。

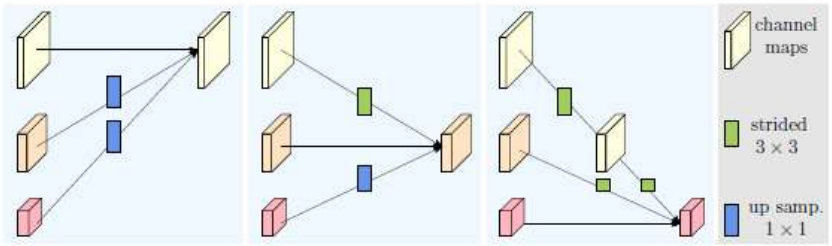

图 2-2：重复多分辨率融合示意图

不同的版本的 HRNet 框架结构拥有不同形式的表达头，其示例图如图 2-3 所示。三种表达头的区别在于输出结果的方式与形式不同。HRNetV1 框架只输出来自高分辨率卷积流的输出表达；HRNetV2 框架中，低分辨率卷积流的输出特征被上采样至高分辨率卷积流的分辨率，随后该框架将四个分辨率卷积流的输出结果聚合，并通过 1*1 的卷积将这四种输出特征混合；HRNetV2p 框架中，在 HRNetV2 框架的基础上，以特征金字塔的形式输出 HRNetV2 框架的结果特征。在本论文的实验中，三种框架被用在不同应用的实验中进行验证测试，HRNetV1 用于人类姿态检测、HRNetV2 用于图像分割、HRNetV2p 用于目标检测。

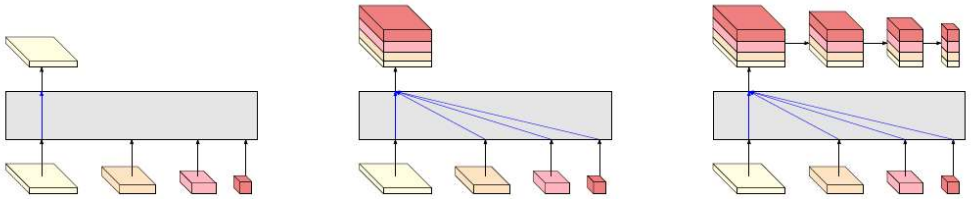

图 2-3：不同 HRNet 表达头示例图，从左到右分别为 HRNetV1、HRNetV2、HRNetV2p

本论文提供了 HRNet 框架的一个整体示例图，如图 2-4 所示。在现阶段设计的 HRNet 框架中，包含四个阶段，同时四个卷积流的分辨率分别为$1/4$ 、$1/8$、$1/16$和$1/32$。每个阶段由调制化模块组成，并在四个阶段中分别重复 1 次、1 次、4 次和 3 次。在四个阶段中，调制化模块分别由 1 个、2 个、3 个和 4 个分支组成。每个分支对应着不同的分辨率，均由四个残差单元组成，并在分支末尾有一个多分辨率融合单元，用于该阶段的重复多分辨率融合。在表格中，C代表每个残差单元中通道的数目。

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×				$\begin{bmatrix} 3\times3, 8C \\ 3\times3, 8C \end{bmatrix} \times 4 \times 3$

图 2-4：HRNet 框架示例图

3 实验分析

本论文实验采用三个版本的 HRNet 网络框架，分别在三个不同的图像处理应用中进行实验，HRNetV1 用于人类姿态检测、HRNetV2 用于图像分割、HRNetV2p 用于目标检测。

在人类姿态检测实验中，采用 HRNetV1 作为网络框架，平方根误差作为损失函数，带有两像素标准差的二维高斯分布用于生成地面真实标签的热图。同时，在实验中，使用 COCO 数据集作为训练与测试数据集，使用目标关键点相似度作为标准评价指标，在此基础上计算标准平均精度和召回得分，并且在训练过程中采用了多种数据增强措施。实验在测试集上的测试结果如图 3-1 所示，能够看出 HRNetV1 的检测精度明显优于当前流行的其他人类姿态检测算法，以 HRNetV1-W32 为例，该网络不仅实现了精度上的提高，AP 得到 74.9，优于其他算法，同时拥有更小的模型尺寸和更少的计算复杂度。在此基础上，本论文提出了更大的模型 HRNetV1-W48，该模型实现了更高的计算精度，同时，使用了来自 AI Challenger 的额外数据进行训练，精度得到了进一步提升，AP 得分达到了最高的 77.0。这展示了 HRNetV1 网络框架在进行人类姿态检测的应用中的优越性，该框架明显优于流行的自底向上的算法，并在模型尺寸、计算复杂度方面更有优势，当扩大网络结构、与其他算法采用相同设置时，检测精度更高。

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**

图 3-1：HRNetV1 在 COCO 测试集上的测试结果

在图像分割检测实验中，采用 HRNetV2 作为网络框架，softmax 作为损失函数，平均分级类别交并比作为验证指标。同时，在本实验中采用三个数据集进行实验，分别为两个场景解析数据集 PASCAL-Context、Cityscapes 和一个人体解析数据集 LIP，并且在训练过程中采用了数据增强策略、随机梯度下降优化器、动态调整学习率策略。实验在 Cityscapes 测试集上的结果如图 3-2 所示，在 PASCAL-Context 测试集上的结果如图 3-3 所示，在 LIP 测试集上的结果如图 3-4 所示。

TABLE 4
Semantic segmentation results on Cityscapes test. We use HRNetV2-W48, whose parameter complexity and computation complexity are comparable to dilated-ResNet-101 based networks, for comparison. Our results are superior in terms of the four evaluation metrics. The result from the combination with OCR [170] is further improved. D-ResNet-101 = Dilated-ResNet-101.

	backbone	mIoU	iIoU cla.	IoU cat.	iIoU cat.
Model learned on the train set					
PSPNet [181]	D-ResNet-101	78.4	56.7	90.6	78.6
PSANet [182]	D-ResNet-101	78.6	-	-	-
PAN [82]	D-ResNet-101	78.6	-	-	-
AAF [69]	D-ResNet-101	79.1	-	-	-
HRNetV2	HRNetV2-W48	**80.4**	**59.2**	**91.5**	**80.8**
Model learned on the train+val set					
GridNet [42]	-	69.5	44.1	87.9	71.1
LRR-4x [46]	-	69.7	48.0	88.2	74.7
DeepLab [19]	D-ResNet-101	70.4	42.6	86.4	67.7
LC [84]	-	71.1	-	-	-
Piecewise [91]	VGG-16	71.6	51.7	87.3	74.1
FRRN [114]	-	71.8	45.5	88.9	75.1
RefineNet [90]	ResNet-101	73.6	47.2	87.9	70.6
PEARL [65]	D-ResNet-101	75.4	51.6	89.2	75.1
DSSPN [88]	D-ResNet-101	76.6	56.2	89.6	77.8
LKM [111]	ResNet-152	76.9	-	-	-
DUC-HDC [144]	-	77.6	53.6	90.1	75.2
SAC [176]	D-ResNet-101	78.1	-	-	-
DepthSeg [73]	D-ResNet-101	78.2	-	-	-
ResNet38 [151]	WResNet-38	78.4	59.1	90.9	78.1
BiSeNet [166]	ResNet-101	78.9	-	-	-
DFN [167]	ResNet-101	79.3	-	-	-
PSANet [182]	D-ResNet-101	80.1	-	-	-
PADNet [159]	D-ResNet-101	80.3	58.8	90.8	78.5
CFNet [173]	D-ResNet-101	79.6	-	-	-
Auto-DeepLab [95]	-	80.4	-	-	-
DenseASPP [181]	WDenseNet-161	80.6	59.1	90.9	78.1
SVCNet [33]	ResNet-101	81.0	-	-	-
ANN [195]	D-ResNet-101	81.3	-	-	-
CCNet [61]	D-ResNet-101	81.4	-	-	-
DANet [43]	D-ResNet-101	81.5	-	-	-
HRNetV2	HRNetV2-W48	81.6	**61.8**	**92.1**	**82.2**
HRNetV2 + OCR [170]	HRNetV2-W48	**82.5**	61.7	**92.1**	81.6

图 3-2：HRNetV2 在 Cityscapes 测试集上的测试结果

TABLE 5
Semantic segmentation results on PASCAL-Context. The methods are evaluated on 59 classes and 60 classes. Our approach performs the best for 60 classes, and performs worse for 59 classes than APCN [51] that developed a strong contextual method. Our approach, combined with OCR [170], achieves significant gain, and performs the best. D-ResNet-101 = Dilated-ResNet-101.

	backbone	mIoU (59)	mIoU (60)
FCN-8s [125]	VGG-16	-	35.1
BoxSup [29]	-	-	40.5
HO_CRF [2]	-	-	41.3
Piecewise [91]	VGG-16	-	43.3
DeepLab-v2 [19]	D-ResNet-101	-	45.7
RefineNet [90]	ResNet-152	-	47.3
UNet++ [189]	ResNet-101	47.7	-
PSPNet [181]	D-ResNet-101	47.8	-
Ding et al. [32]	ResNet-101	51.6	-
EncNet [172]	D-ResNet-101	52.6	-
DANet [43]	D-ResNet-101	52.6	-
ANN [195]	D-ResNet-101	52.8	-
SVCNet [33]	ResNet-101	53.2	-
CFNet [173]	D-ResNet-101	54.0	-
APCN [51]	D-ResNet-101	55.6	-
HRNetV2	HRNetV2-W48	54.0	48.3
HRNetV2 + OCR [170]	HRNetV2-W48	**56.2**	**50.1**

图 3-3：HRNetV2 在 PASCAL-Context 测试集上的测试结果

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**

图 3-4：HRNetV2 在 LIP 测试集上的测试结果

从 HRNetV2 在 Cityscapes 上的测试结果能够看出，相比于如 UNet++或 DeepLabv3+等模型，HRNetV2 不仅拥有更小的模型尺寸、更低的模型复杂度，同时拥有更高的检测精度，例如 HRNetV2 相比 UNet++模型，精度提升了 5.6 点；从 HRNetV2 在 PASCAL-Context 上的测试结果能够看出，在两个指标上，HRNetV2 模型的检测性能优于其他大部分模型，当配合 OCR 主题的情况下，HRNetV2 的检测性能能够进一步提升；从 HRNetV2 在 LIP 测试集上的测试结果能够看出，HRNetV2-W48 模型仅仅使用更少的模型参数和更小的计算复杂度，但是实现了在整体精度上优于其他模型，更值得一提的是，该模型没有使用任何额外的诸如边缘或姿态的信息。

在目标检测实验中，采用 HRNetV2p 作为模型，在 Faster-RCNN、Cascade R-CNN、FCOS 和 CenterNet 等框架中实现，主要与他们带有 ResNet 或 ResNeXt 模型的标准框架进行对比。实验的训练与测试均在 MMDetection 平台上实现，实验测试结果如图 3-5 所示。从实验结果能够看出，在 Faster-RCNN 框架中，HRNetV2p 的检测效果优于 ResNet，同时与其具有相似的模型尺寸和模型复杂度，类似地与在其他框架中的比较也能够看出，HRNetV2p 展现出更

加优秀的检测效果。

TABLE 11
Comparison with the state-of-the-art single-model object detectors on COCO test-dev with BN parameters fixed and without mutli-scale training and testing. * means that the result is from the original paper [12]. GFLOPs and #parameters of the models are given in Table 7. The observations are similar to those on COCO val, and show that the HRNet performs better than ResNet and ResNeXt under state-of-the-art object detection and instance segmentation frameworks.

	backbone	size	LS	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
MLKP [142]	VGG16	-	-	28.6	52.4	31.6	10.8	33.4	45.1
STDN [187]	DenseNet-169	513	-	31.8	51.0	33.6	14.4	36.1	43.4
DES [179]	VGG16	512	-	32.8	53.2	34.6	13.9	36.0	47.6
CoupleNet [194]	ResNet-101	-	-	33.1	53.5	35.4	11.6	36.3	50.1
DeNet [139]	ResNet-101	512	-	33.8	53.4	36.1	12.3	36.1	50.8
RFBNet [96]	VGG16	512	-	34.4	55.7	36.4	17.6	37.0	47.6
DFPR [74]	ResNet-101	512	1×	34.6	54.3	37.3	-	-	-
PFPNet [70]	VGG16	512	-	35.2	57.6	37.9	18.7	38.6	45.9
RefineDet [172]	ResNet-101	512	-	36.4	57.5	39.5	16.6	39.9	51.4
Relation Net [56]	ResNet-101	600	-	39.0	58.6	42.9	-	-	-
C-FRCNN [25]	ResNet-101	800	1×	39.0	59.7	42.8	19.4	42.4	53.0
RetinaNet [93]	ResNet-101-FPN	800	1.5×	39.1	59.1	42.3	21.8	42.7	50.2
Deep Regionlets [160]	ResNet-101	800	1.5×	39.3	59.8	-	21.7	43.7	50.9
FitnessNMS [140]	ResNet-101	768	-	39.5	58.0	42.6	18.9	43.5	54.1
DetNet [86]	DetNet59-FPN	800	2×	40.3	62.1	43.8	23.6	42.6	50.0
CornerNet [79]	Hourglass-104	511	-	40.5	56.5	43.1	19.4	42.7	53.9
M2Det [185]	VGG16	800	~ 10×	41.0	59.7	45.0	22.1	46.5	53.8
Faster R-CNN [92]	ResNet-101-FPN	800	1×	39.3	61.3	42.7	22.1	42.1	49.7
Faster R-CNN	HRNetV2p-W32	800	1×	39.5	61.2	43.0	23.3	41.7	49.1
Faster R-CNN [92]	ResNet-101-FPN	800	2×	40.3	61.8	43.9	22.6	43.1	51.0
Faster R-CNN	HRNetV2p-W32	800	2×	41.1	62.3	44.9	24.0	43.1	51.4
Faster R-CNN [92]	ResNet-152-FPN	800	2×	40.6	62.1	44.3	22.6	43.4	52.0
Faster R-CNN	HRNetV2p-W40	800	2×	42.1	63.2	46.1	24.6	44.5	52.6
Faster R-CNN [17]	X-101-64×4d-FPN	800	2×	41.1	62.8	44.8	23.5	44.1	52.3
Faster R-CNN	HRNetV2p-W48	800	2×	42.4	63.6	46.4	24.9	44.6	53.0
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
FCOS [136]	ResNet-50-FPN	800	2×	37.3	56.4	39.7	20.4	39.6	47.5
FCOS	HRNetV2p-W18	800	2×	37.8	56.1	40.4	21.6	39.8	47.4
FCOS [136]	ResNet-101-FPN	800	2×	39.2	58.8	41.6	21.8	41.7	50.0
FCOS	HRNetV2p-W32	800	2×	40.5	59.3	43.3	23.4	42.6	51.0
CenterNet [36]	Hourglass-52	511	–	41.6	59.4	44.2	22.5	43.1	54.1
CenterNet	HRNetV2-W48	511	–	43.5	62.1	46.5	22.2	46.5	57.8
Cascade Mask R-CNN [13]	ResNet-101-FPN	800	~ 1.6×	44.0	62.3	47.9	24.3	46.9	56.7
Cascade Mask R-CNN	HRNetV2p-W32	800	~ 1.6×	44.7	62.5	48.6	25.8	47.1	56.3
Cascade Mask R-CNN [13]	X-101-64 × 4d-FPN	800	~ 1.6×	45.9	64.5	50.0	26.6	49.0	58.6
Cascade Mask R-CNN	HRNetV2p-W48	800	~ 1.6×	46.1	64.0	50.3	27.1	48.6	58.3
Hybrid Task Cascade [16]	ResNet-101-FPN	800	~ 1.6×	45.1	64.3	49.0	25.2	48.0	58.2
Hybrid Task Cascade	HRNetV2p-W32	800	~ 1.6×	45.6	64.1	49.4	26.7	47.7	58.0
Hybrid Task Cascade [16]	X-101-64 × 4d-FPN	800	~ 1.6×	47.2	66.5	51.4	27.7	50.1	60.3
Hybrid Task Cascade	HRNetV2p-W48	800	~ 1.6×	47.0	65.8	51.0	27.9	49.4	59.7
Hybrid Task Cascade [16]	X-101-64 × 4d-FPN	800	~ 2.3×	47.2	66.6	51.3	27.5	50.1	60.6
Hybrid Task Cascade	HRNetV2p-W48	800	~ 2.3×	47.3	65.9	51.2	28.0	49.7	59.8

图 3-5：HRNetV2p 在 COCO 测试集上的测试结果

4 消融实验

本论文通过消融实验的方式，对 HRNet 框架的参数进行了对比分析，主要在不同分辨率表达、重复分辨率融合、分辨率维护、不同版本的 HRNet 四个方面进行分析。

在不同分辨率表达的实验中，通过检查从高到低分辨率输出预测热图的质量，比较分辨率表达如何影响姿态检测的表现。实验结果如图 4-1 所示，从实验结果能够看出，高分辨率特征的检测精度高于中分辨率与低分辨率，这暗含了高分辨率的确促进了检测表现。

在重复分辨率融合的实验中，研究不同融合方式对检测结果的影响，在实验中分别控制网络在最终输出时融合、在处理的每个阶段末尾时融合、在每个阶段交叉融合。实验结果如图 4-2 所示，从实验结果能够看出，在三个使其均使用分辨率融合的网络检测精度最佳，基

于此能够得出，多分辨率融合单元对检测精度的提升是有益的，同时更多的融合将带来更加优秀的检测效果。

在分辨率维护方面的实验中，论文提出了 HRNet 框架模型的一种变体，该变体将四个从高到低的卷积流贯穿于网络始终，同时这四个卷积流拥有相同的深度。通过在人类姿态检测与图像分割方面的实验能够看出，该变体网络相较于论文提出的 HRNet 具有较低的检测性能，作者认为这是因为在网络处理的早期阶段，低分辨率卷积流提取到的低分辨率特征对于最终检测精度的提升是少有帮助的。类似地，作者采用了另一种只有高分辨率卷积流的 HRNet 变体进行实验，发现这种变体的检测精度同样低于论文提出的 HRNet 框架模型，这说明分阶段将从高到低分辨率的卷积流进行并行卷积是十分必要的。

在不同版本的 HRNet 模型实验中，本论文将 HRNetV1、HRNetV1 的变体、HRNetV2 框架分别在图像分割和目标检测任务上进行实验，其中该变体是在 HRNetV1 后附加 1*1 的卷积，使得输出高分辨率尺寸维度与 HRNetV2 一致，目标检测任务中使用的是 HRNetV2p 而不是 HRNetV2。实验结果如图 4-3 所示。从先前的关于人类姿态检测的实验中能够看出，HRNetV1 与 HRNetV2 的检测效果是相似的。从本次消融实验的实验结果能够看出，HRNetV2 在两个视觉任务中的检测效果均是优于 HRNetV1 与 HRNetV1 变体模型的。这表明了如同 HRNetV2 版本的框架结构在输出表达中聚合低分辨率并行卷积流的特征是基本的提高网络检测能力的步骤。

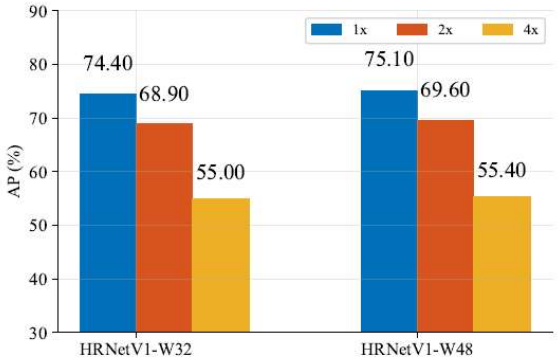

Fig. 9. Ablation study about the resolutions of the representations for human pose estimation. 1×, 2×, 4× correspond to the representations of the high, medium, low resolutions, respectively. The results imply that higher resolution improves the performance.

图 4-1：不同分辨率表达的消融实验结果

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

图 4-2：重复分辨率融合的消融实验结果

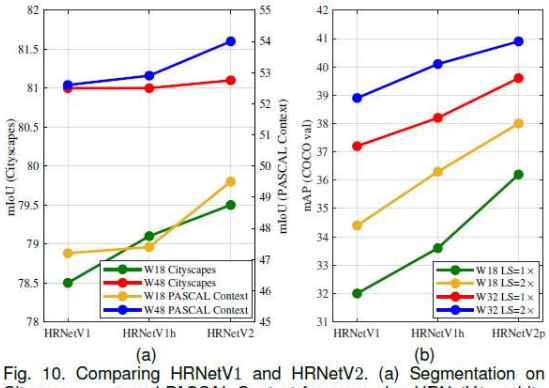

Fig. 10. Comparing HRNetV1 and HRNetV2. (a) Segmentation on Cityscapes `val` and PASCAL-Context for comparing HRNetV1 and its variant HRNetV1h, and HRNetV2 (single scale and no flipping). (b) Object detection on COCO `val` for comparing HRNetV1 and its variant HRNetV1h, and HRNetV2p (LS = learning schedule).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

图 4-3：不同版本的 HRNet 模型消融实验结果

5 实验结论与展望

本论文提出了一种新型的轻量级高分辨率网络，解决当前的视觉识别问题。该网络的创新点主要体现在三个方面：第一，通过并行而非序列的方式连接高分辨率和低分辨率卷积流；第二，在整个处理过程中维持图像的高分辨率表达，而不是将最终高分辨率特征输出结果从低分辨率中恢复；第三，不断进行多分辨率融合，使高分辨率表达具有更强的位置敏感性。通过本论文的工作，作者希望的是科研人员能够将研究重心向通过直接设计网络结构解决特定视觉问题方面转移，而不是将输出结果从低分辨率特征中扩展、修复得来。

本论文对关于 HRNet 未来的研究方向进行了展望。论文指出，关于 HRNet 在图像分割领域中的研究，能够与其他如 OCR 等技术相结合，提高图像分割或实例分割的精度。另一方面，应当尽力挖掘 HRNet 的潜力，使其应用在其他更多与位置相关的视觉应用中，如无人机目标检测、图像风格化和图像增强等领域。