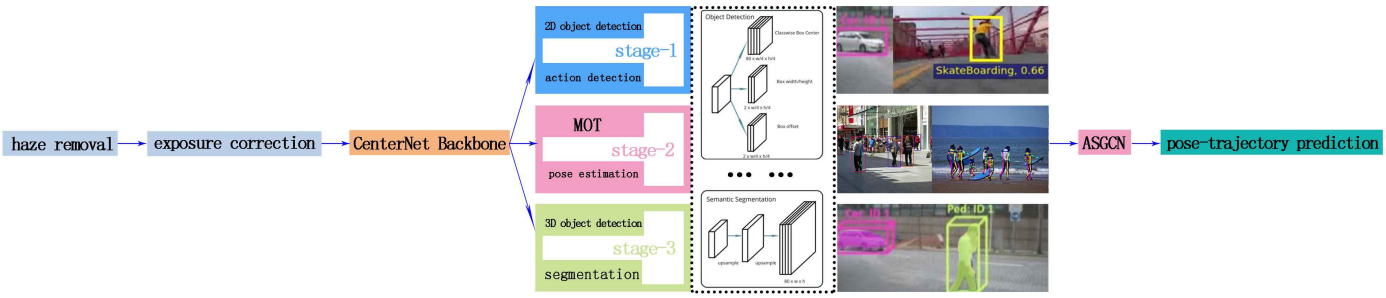
项目网络架构



① 背景建模简化——固定摄像头

② 空间特征尺度——中小尺度人体

③ 多任务学习

④ 边云协同

CenterNet与多任务学习(MTL)

CenterNet主要思想

基于从物体中心点提取的特征进行回归任务，实现2D目标检测、3D目标检测、姿态估计（人体关键点检测）、多目标追踪、图像分割、时空动作检测（输入一段视频，识别动作出现的区间和对应的类别，并用一个框出人物的空间位置）

多任务学习（MTL）

硬约束

不同任务通过分享一些底部的层或backbone网络学习一些可共享的特征信息，为了保证任务的独特性，每个任务在顶部拥有自己独特的层学习各自的特征。

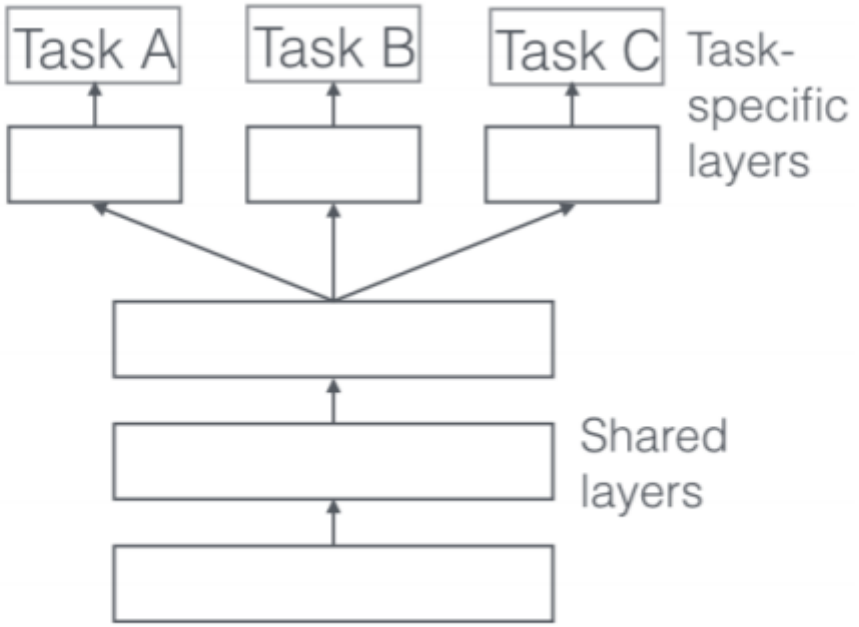


Figure 1: Hard parameter sharing for multi-task learning in deep neural networks

软约束

隐藏层参数共享，不同的任务可以使用不同的网络，但是不同任务的网络参数，采用正则化作为约束，底层的参数不一定完全一致，只是鼓励参数相似化。

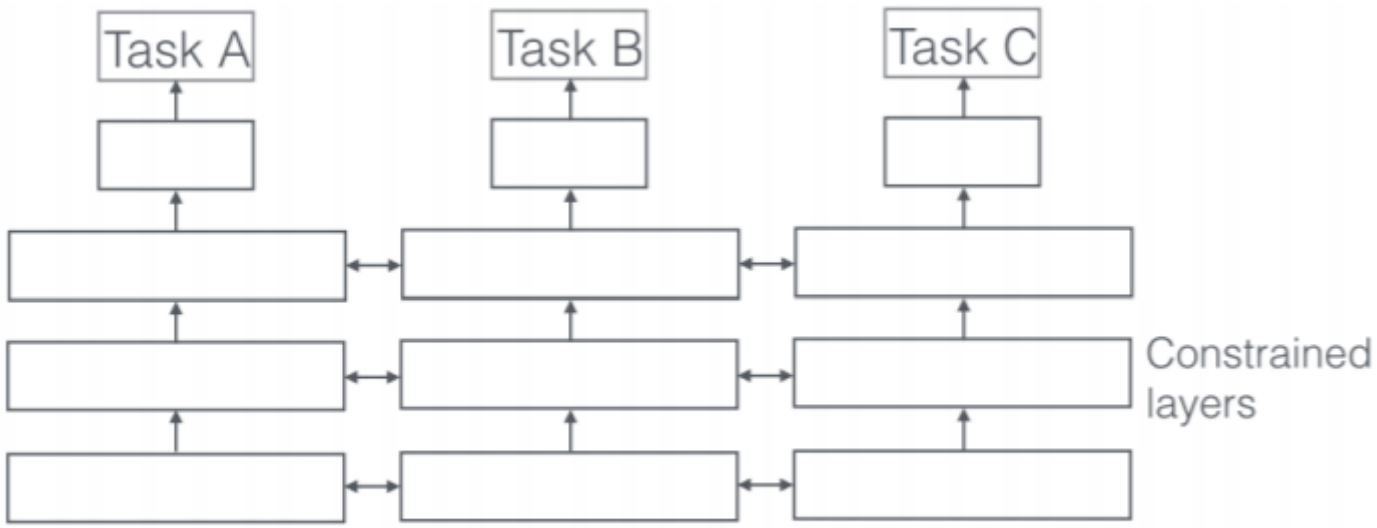


Figure 2: Soft parameter sharing for multi-task learning in deep neural networks

总之，多任务学习的关键就在于寻找任务之间的关系，如果任务之间的关系衡量恰当，那么不同任务之间就能相互提供额外的有用信息，利用这些额外信息，可以训练出表现更好、更鲁棒的模型。反之，如果关系衡量不恰当，不仅不会引入额外的信息，反而会给任务本身引来噪声，模型学习效果不升反降。

CenterNet简述

网络输入为宽W高h的输入图像

$$I \in R^{W \times H \times 3}$$

目标输出为关键点热图，R为输出步长，输入图像以输出因子R被下采样 （默认R = 4），C为关键点类型数，在目标检测中C = 80个目标类别，这是因为官方数据集有80个类，在检测人体关键点时，C可以是人体关键点数目。

$$\hat{Y} \in [0,1]^{\frac{W}{R} \times \frac{H}{R} \times C}$$

在关键点热图中该点的值为1，代表检测到的一个关键点

$$\hat{Y}_{x,y,c} = 1$$

在关键点热图中该点的值为0，代表检测到的是背景

$$\hat{Y}_{x,y,c} = 0$$

将所有的GT关键点使用一个高斯核映射到GT热图Y

$$Y_{xyc} = \exp\left(-\frac{(x-\tilde{p}_x)^2+(y-\tilde{p}_y)^2}{2\sigma_p^2}\right)$$

$$Y \in [0,1]^{\frac{W}{R} \times \frac{H}{R} \times C}$$

为了恢复输出步长引起的离散化误差，对每个中心点预测一个局部偏移

$$\hat{O} \in \mathcal{R}^{\frac{W}{R} \times \frac{H}{R} \times 2}$$

对所有目标类别使用单个大小预测

$$\hat{S} \in \mathcal{R}^{\frac{W}{R} \times \frac{H}{R} \times 2}$$

计算框四个角点的位置

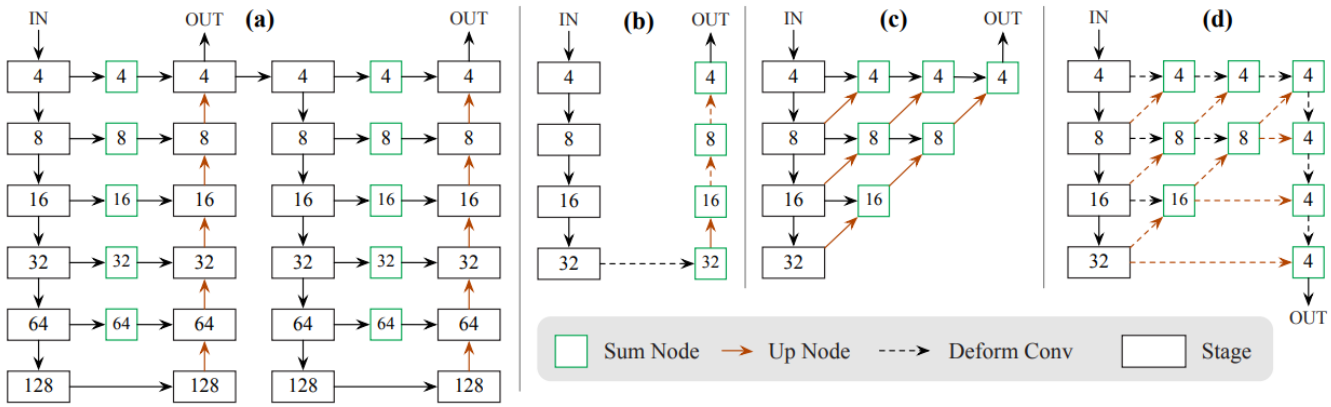
$$\begin{aligned}&(\hat{x}_i + \delta\hat{x}_i - \hat{w}_i/2,\ \ \hat{y}_i + \delta\hat{y}_i - \hat{h}_i/2,\\ &\ \hat{x}_i + \delta\hat{x}_i + \hat{w}_i/2,\ \ \hat{y}_i + \delta\hat{y}_i + \hat{h}_i/2),\end{aligned}$$

$$(\delta\hat{x}_i, \delta\hat{y}_i) \ = \ \hat{O}_{\hat{x}_i, \hat{y}_i}$$

$$(\hat{w}_i, \hat{h}_i) = \hat{S}_{\hat{x}_i, \hat{y}_i}$$

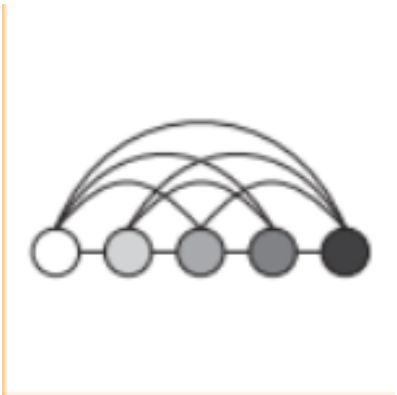
Backbone

Hourglass Network、ResNet、DLA



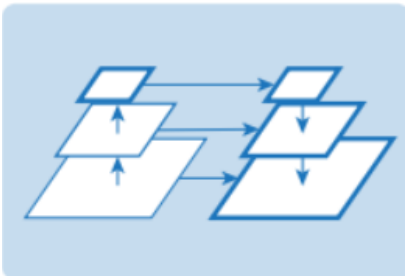
语义融合

浅层的特征信息(re-ID)、中层的特征信息(pose estimation)、深层的特征信息(orientation)



空间融合

上采样、下采样形成不同尺度（分辨率）的特征图，构建特征金字塔，例如：检测人体关键点时，对于小尺度的人体，希望尽可能在大尺度的特征图上进行计算，否则会出现检测失效（openpose的典型问题），对于大尺度的人体，希望在尺度适合的特征图上进行计算即可，以减少计算量。



2D目标检测（火焰、行人、车辆）

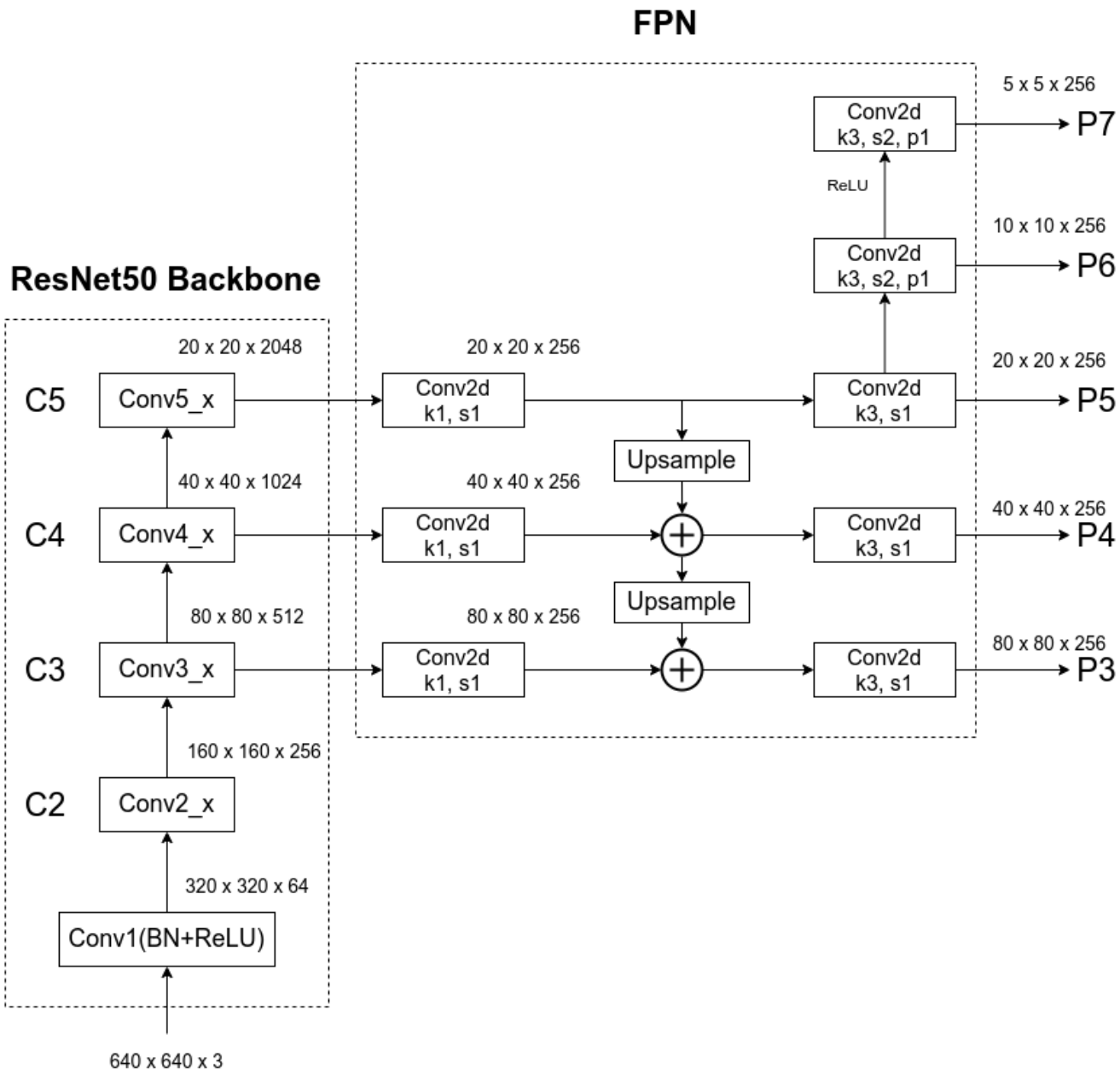
CenterNet（Objects as Points，2019）

CenterNet2 (Probabilistic two-stage detection，2021，SOTA)

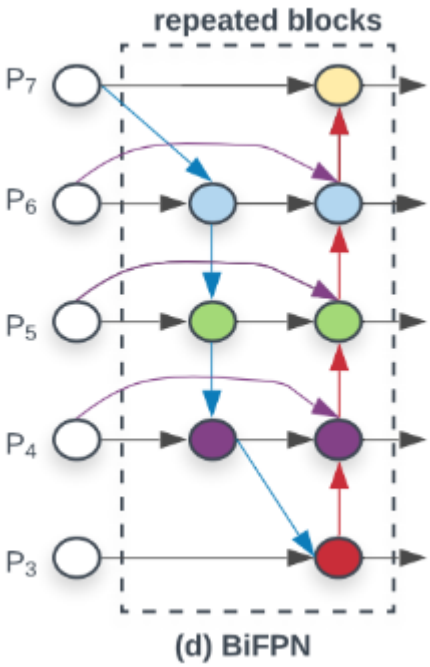
centernet原作者对centernet进行了改进，将一阶段目标检测过程转换为两阶段过程，更快且精度更高，并且给出了更多优化的backbone

FPN

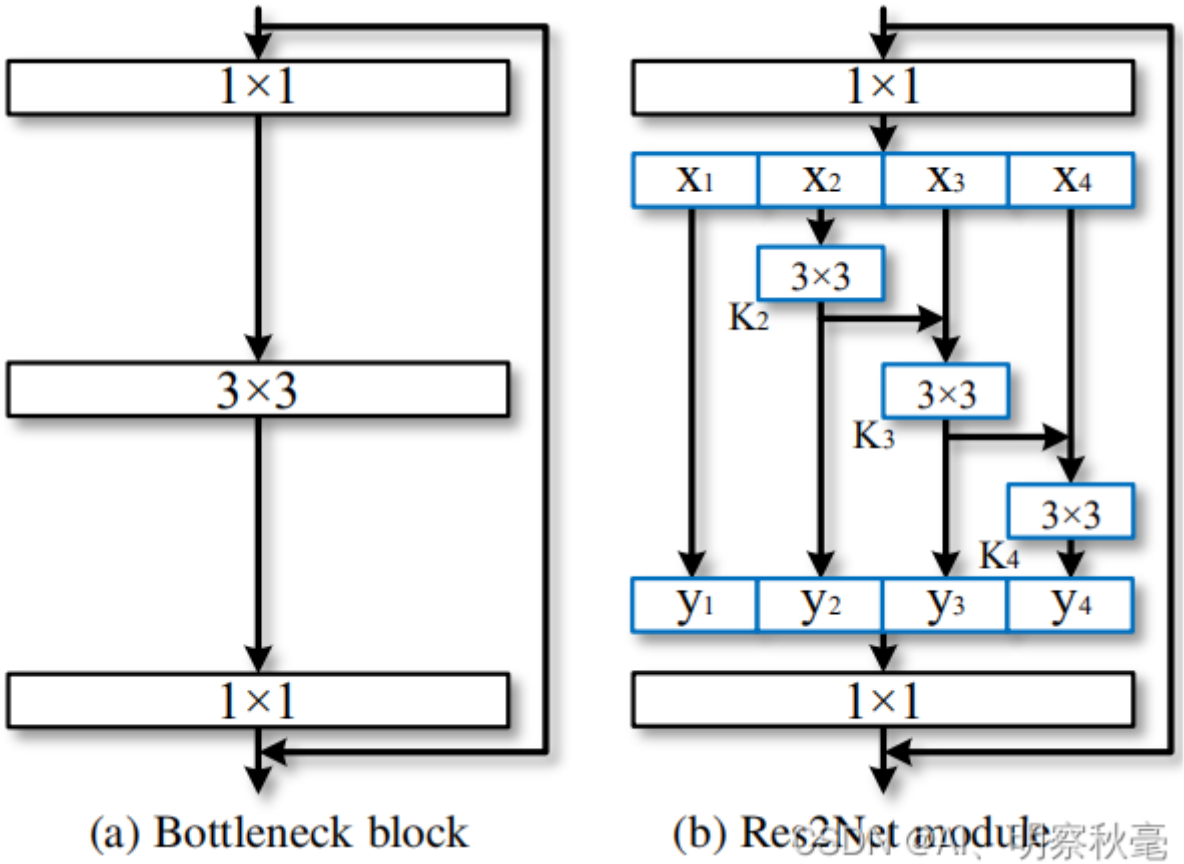
在原resnet上加FPN的方法图示如下



BiFPN



Res2Net



ResNext

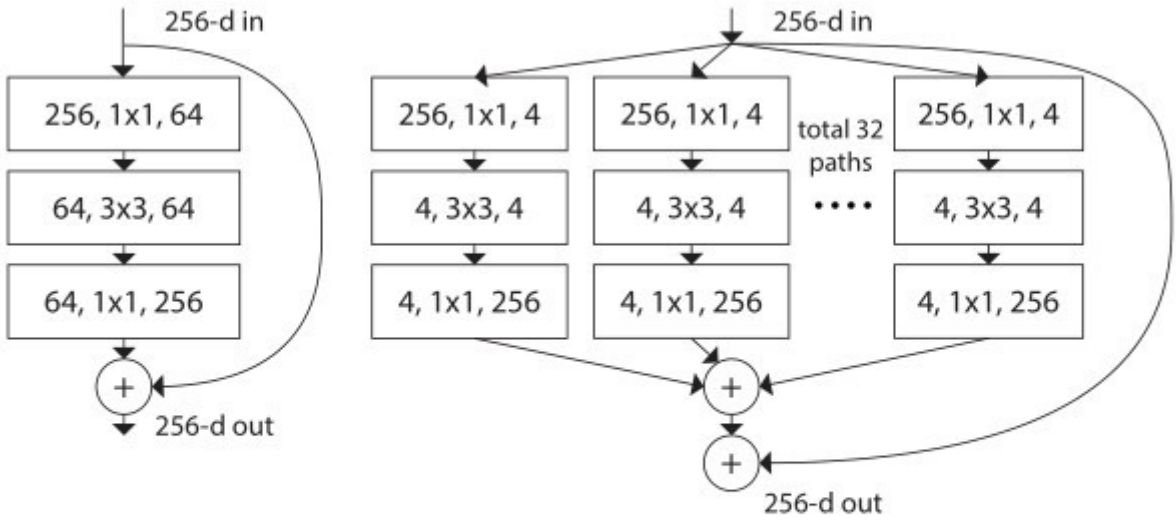


Figure 1. **Left**: A block of ResNet [14]. **Right**: A block of ResNeXt with cardinality = 32, with roughly the same complexity. A layer is shown as (# in channels, filter size, # out channels).

HRNet/HigherHRNet

这些网络结构并不是CenterNet2论文里用过的，而是在人体关键点检测相关的论文里被使用过的，但在网络结构设计上，聚合不同层次、不同尺度的特征信息的这一思路是相通的（事实上，CenterNet论文里的hourglass网络最开始是在人体关键点检测领域的论文里提出的，只要backbone提取的特征满足需求且尽可能好即可，而不用局限于网络的应用领域）

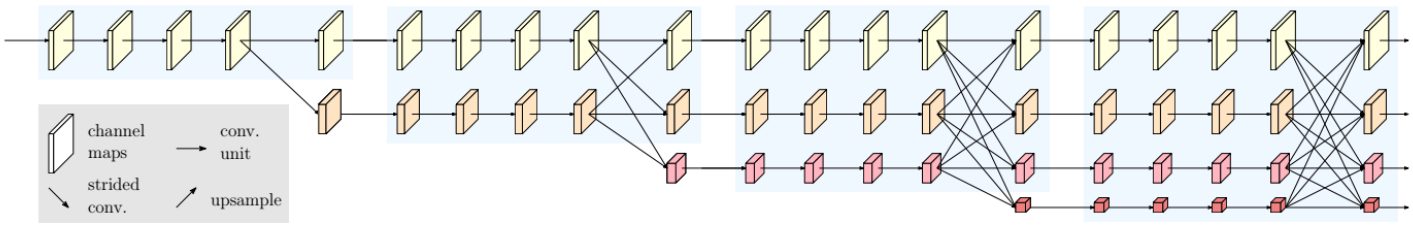


Fig. 2. An example of a high-resolution network. Only the main body is illustrated, and the stem (two stride-2 3×3 convolutions) is not included. There are four stages. The 1st stage consists of high-resolution convolutions. The 2nd (3rd, 4th) stage repeats two-resolution (three-resolution, four-resolution) blocks. The detail is given in Section 3.

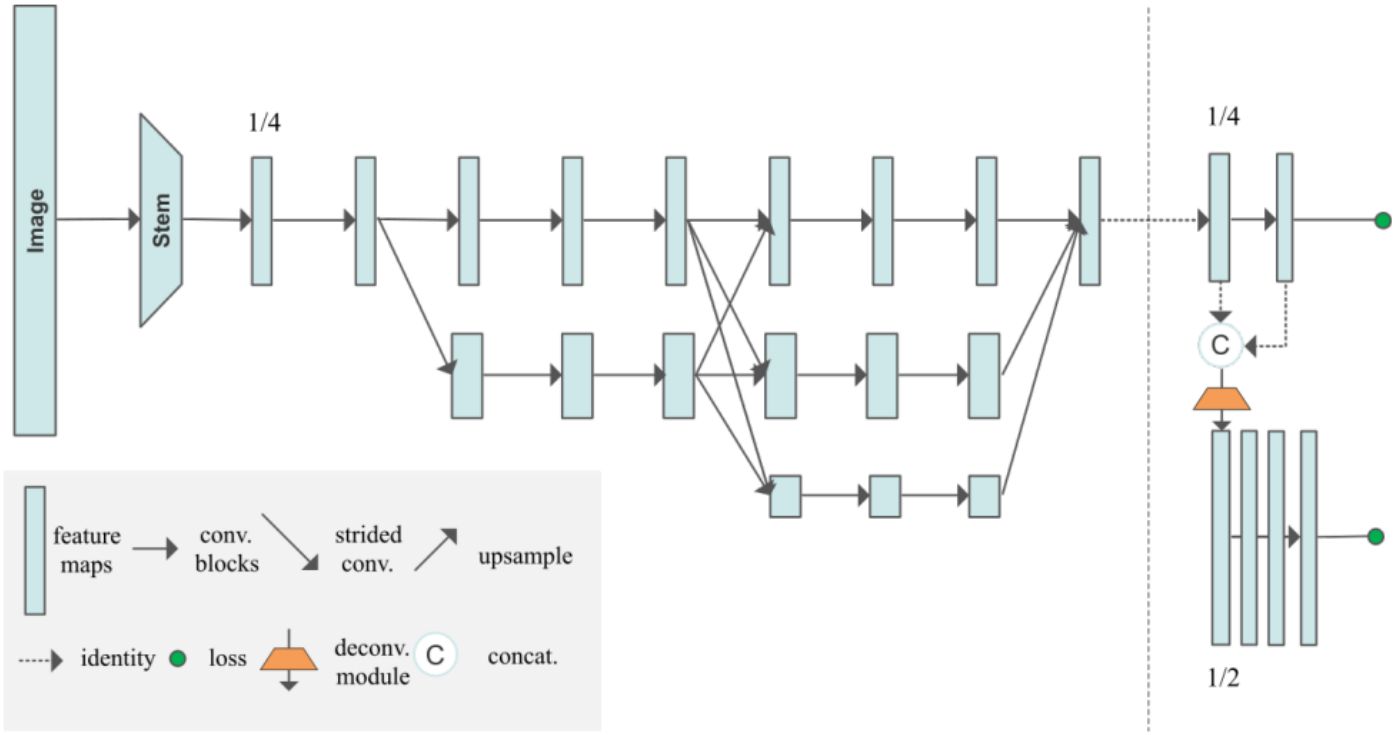


Figure 2. An illustration of HigherHRNet. The network uses HRNet [38, 40] as backbone, followed by one or more deconvolution modules to generate multi-resolution and high-resolution heatmaps. Multi-resolution supervision is used for training. More details are given in Section 3.

时空动作检测

技术分类

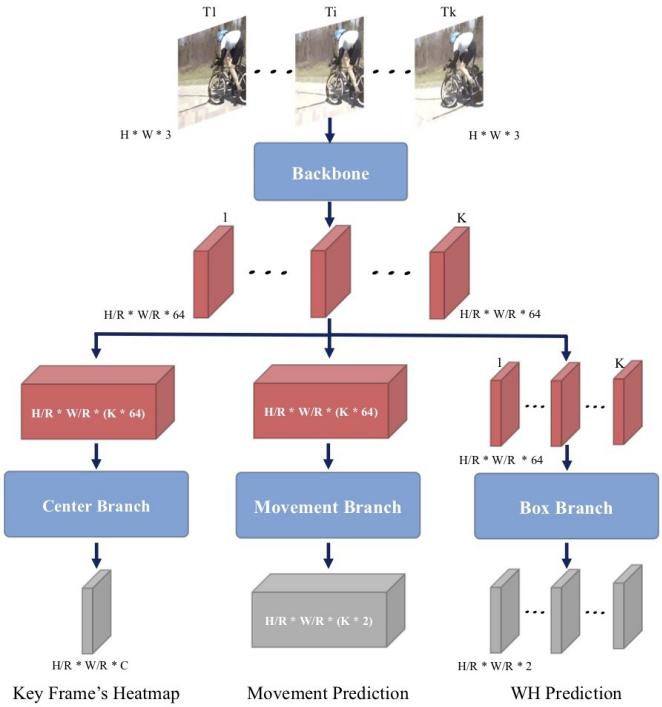
① 动作识别 (action recognition)： 是对每个输入视频进行分类，识别出视频中人物做出的动作。**即输入视频序列，得到视频对应的类别；**

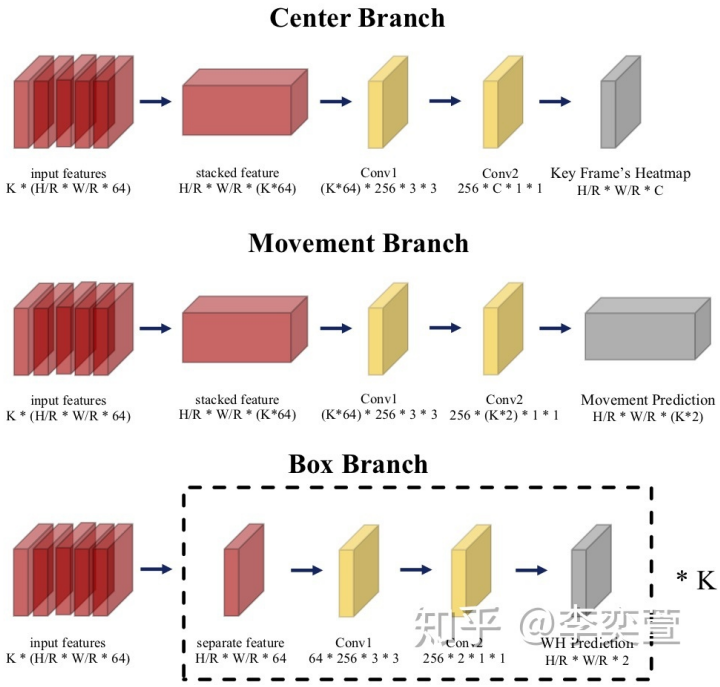
② 时序动作检测 (temporal action detection)： 任务的输入是一个未经裁剪的视频 (untrimmed video)，即在这个视频里有些帧是没有动作发生的，因此需要**检测出动作开始和结束的区间，并判断区间内动作的类别。即输入未经裁剪的视频序列，得到动作出现的区间和对应的类别；**

③ 时空动作检测 (spatio-temporal action detection)： 相比于时序动作检测略有不同，**时空动作检测不仅需要识别动作出现的区间和对应的类别，还要在空间范围内用一个包围框 (bounding box)标记出人物的空间位置。**

Actions as Moving Points (ECCV 2020 SOTA)

将动作实例建模为每一帧动作中心点沿时序的运动轨迹，虽然这样的建模方式直观感觉有些简单粗糙，但是从实验效果来看，确实能够达到不错的准确率，而且从项目需求角度来讲，这样的建模方式应该足够应对摔倒、打架之类不要求细节的动作（比如：花滑是一种要求细节的动作形式，脚的内翻、外翻完全是两种得分不同的动作）。左边是MOC的网络结构，红色的立方体是backbone抽取的feature或者是后期延时序拼接的feature；右边是每个分支的具体结构，每个分支由一个33卷积层，一个ReLu层，一个11卷积层组成，由黄色立方体代表（ReLu被忽略）





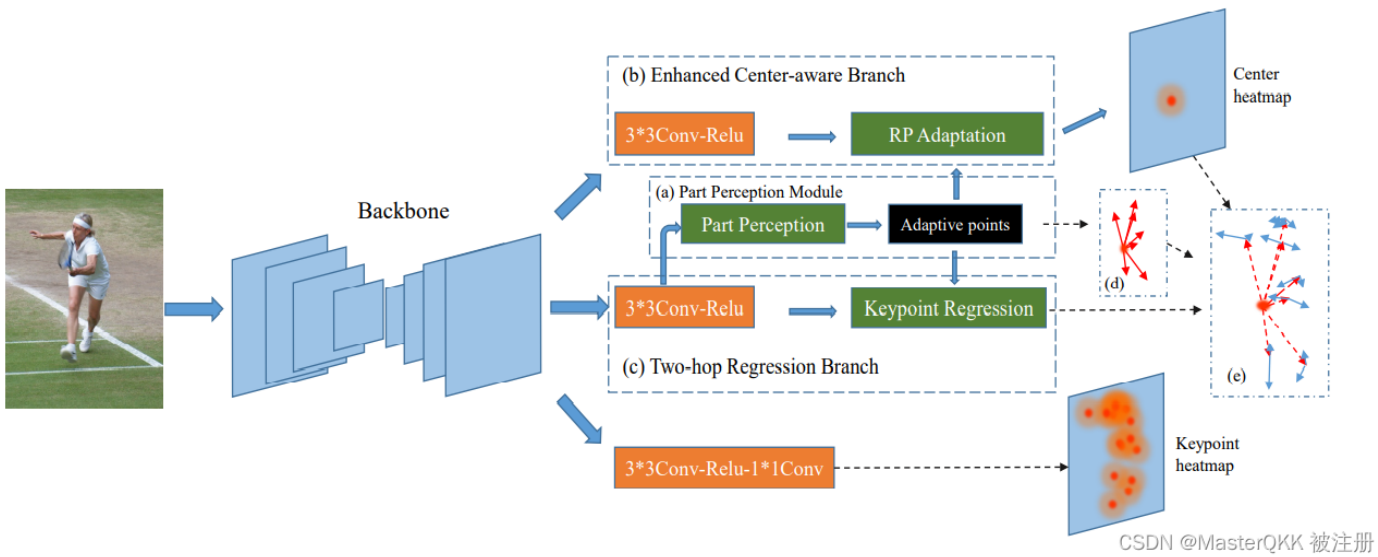
人体关键点检测

CenterNet（Objects as Points，2019）

从中心点直接回归到相应关键点的offset，效果相比SOTA而言并不是很好

AdaptivePose++:（A Powerful Single-Stage Network for Multi-Person Pose Regression，AAAI 2022 SOTA）

通过中心点表示人的实例，并利用中心到关节的偏移量来形成人的姿态, 然而这种方式的问题在于：由于各种姿势的变形和中心的接受场的改变 因此很难处理中心到关节的长距离偏移问题，因此将人体部分表示为自适应点并使用一个自适应点集，包括人类中心和7个与人体部位相关的点来表示不同的人体实例。



① 利用部分感知模块Part Perception Module，从每个人的实例的假定中心回归七个自适应的人体部分相关点；

② 然后，在增强型中心感知分支Enhanced Center-aware Branch中通过聚合自适应点的特征来预测中心热图；

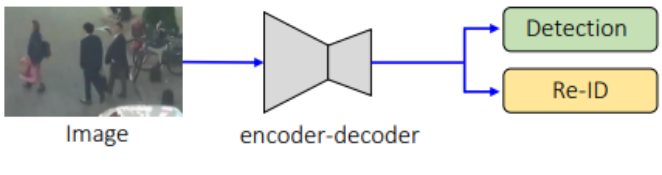
③ 最后，两跳回归分支Two-hop Regression Branch将自适应的人体部分相关点作为中间跳点，间接回归从中心到每个关键点的偏移量，这样也不容易累积误差。

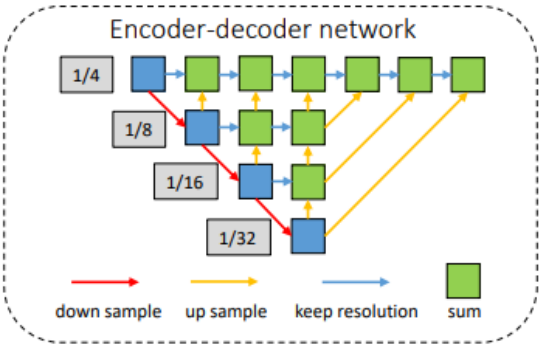
用一些层次比较深的网络作为backbone的话可以达到SOTA水平，直接用DLA34的话，精度只会下降2-3个点，但是速度×2，且adaptivepose对不同尺度的人体自适应了不同尺度的感受野，能够更好地解决校园监控视频中等尺度的人体的姿态估计问题。

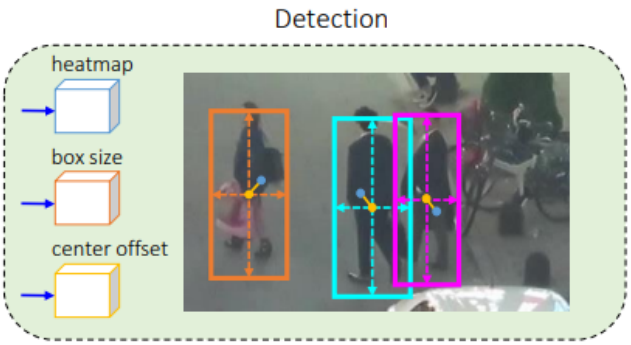
多目标追踪

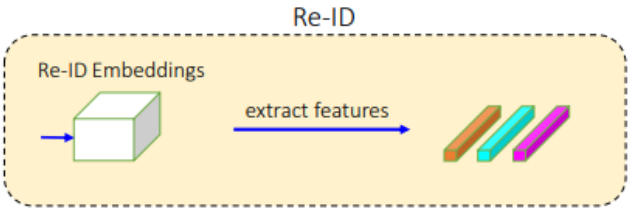
FairMOT（A Simple Baseline for Multi-Object Tracking，CVPR 2020）

在FairMOT发表之后，只出现了少数几篇（目前只找到1篇）论文实现了将准确度提升了2~4个点（但遗憾，代码没开源），但是从准确度和速度的平衡来讲，FairMOT仍然是目前为止最适合的方法，FairMOT至少比其他方法快2倍。









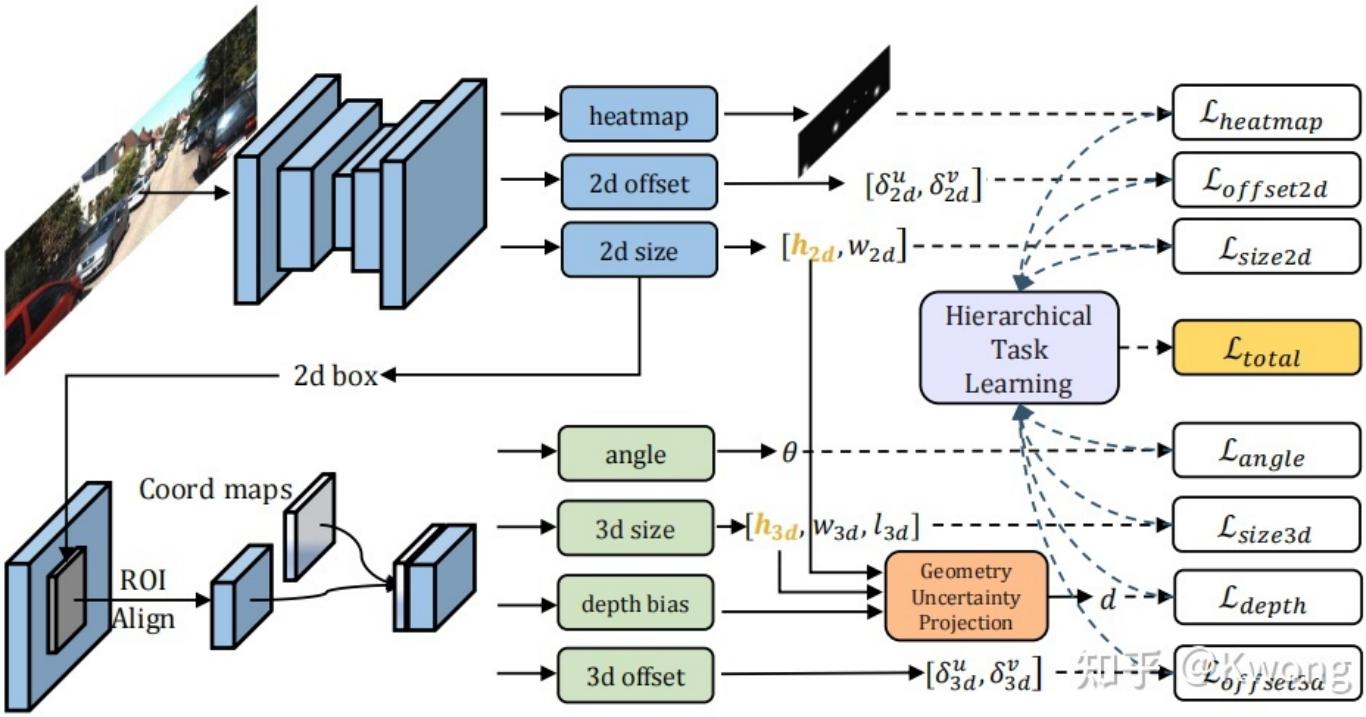
3D目标检测

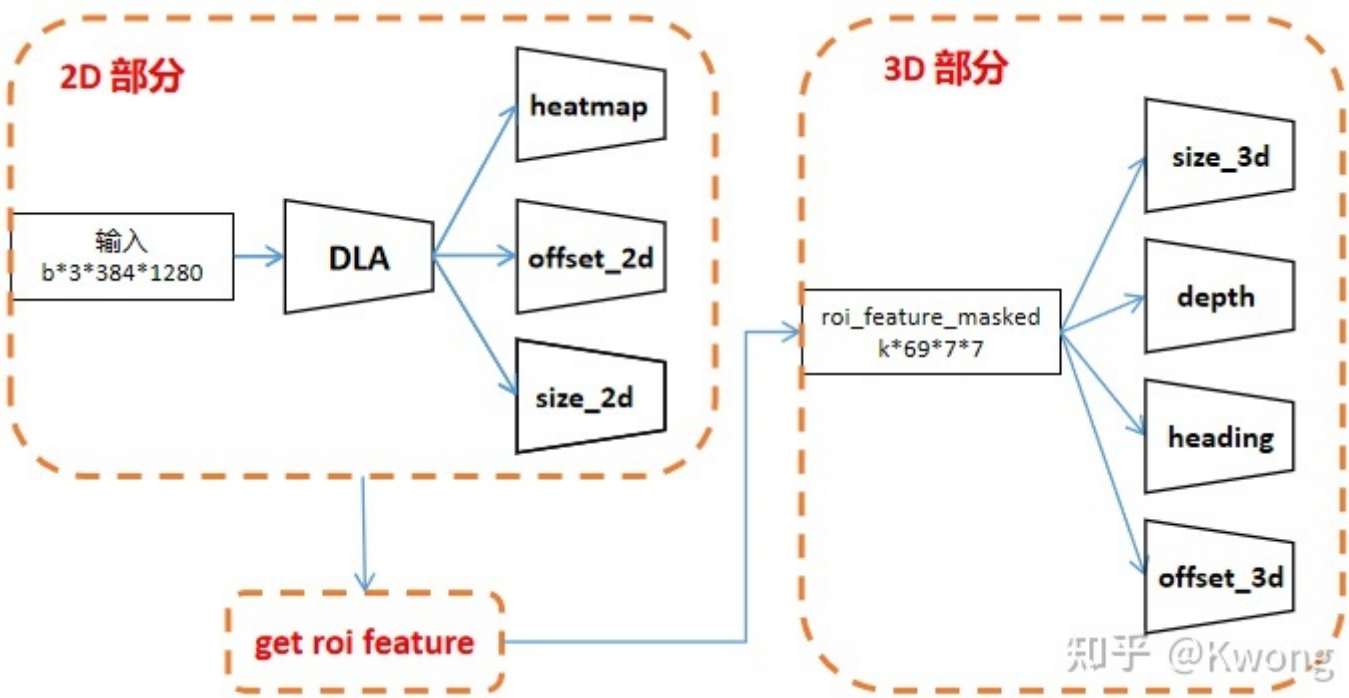
CenterNet（Objects as Points，2019）

相比于2D目标检测而言，3D目标检测需要回归的信息更多一些：深度（目前数据集多以米为单位）、方向（一个角度值）、框从二维变成三维（长宽高），CenterNet原论文中的方法接近于直接从中心点进行回归（在数值变化和损失函数上存在一些细节不同），但是在原论文中，2D目标检测和3D目标检测的head完全是独立的，二者不存在联系，准确度并不是很好（相比SOTA而言）。

GUPNet（Geometry Uncertainty Projection Network for Monocular 3D Object Detection，ICCV 2021 SOTA）

从输入输出流来看，在进行3D物体检测更多地利用了2D物体检测head的信息，而且在对3D物体检测的head进行更进一步的优化设计，减小了误差，提升了推理的可靠性以及训练的稳定性。



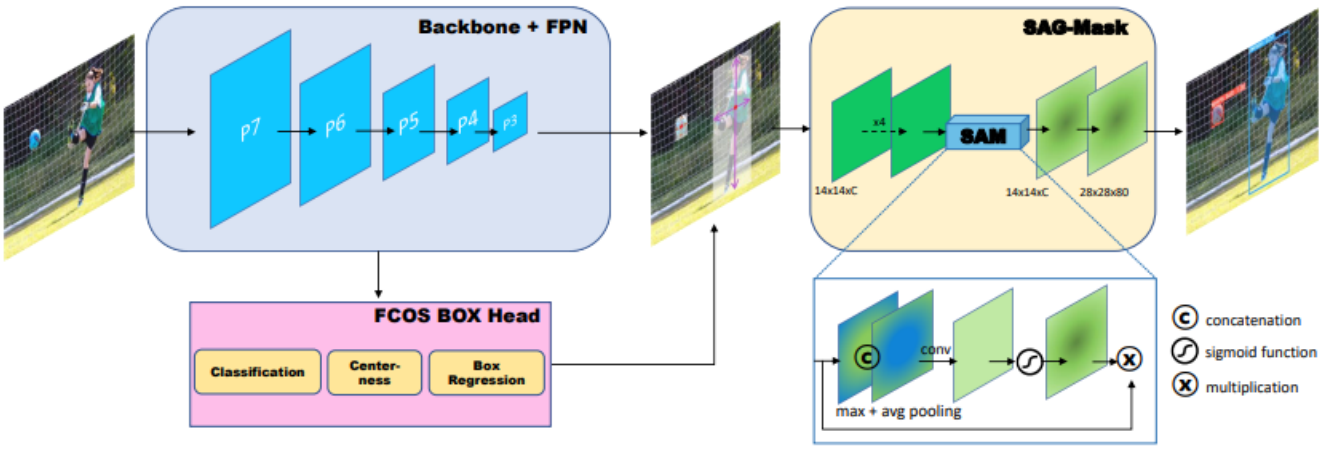


图像分割

关于图像分割，目前项目中未知存在需要用图像分割实现的功能，不过很多文献提到过，在多任务学习中，图像分割和人体姿态关键点检测一起训练能够得到更好的效果。

目前并未对图像分割进行调研，对其原理还并不了解，下面只是列举一篇基于centernet的图像分割论文

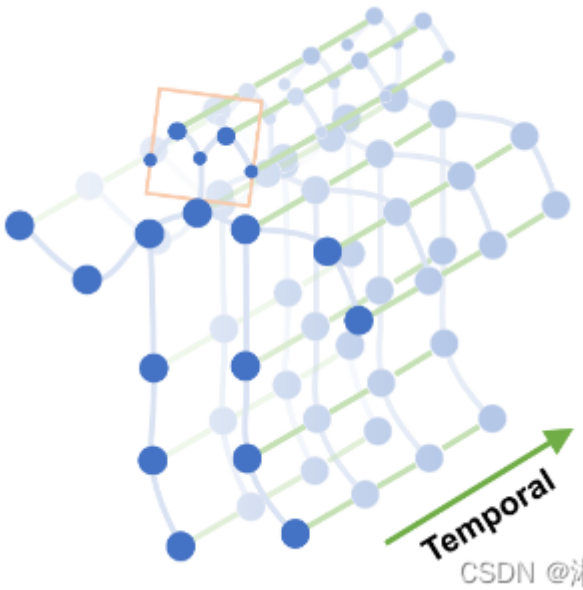
CenterMask（Real-Time Anchor-Free Instance Segmentation，CVPR 2020 SOTA）

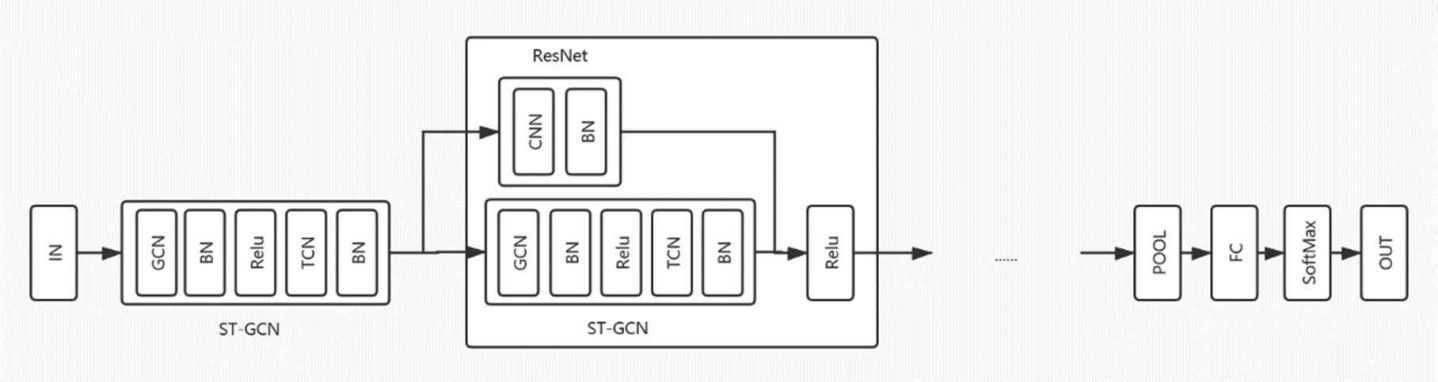


基于人体关键点的动作分类

ST-GCN（Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition，AAAI 2018）

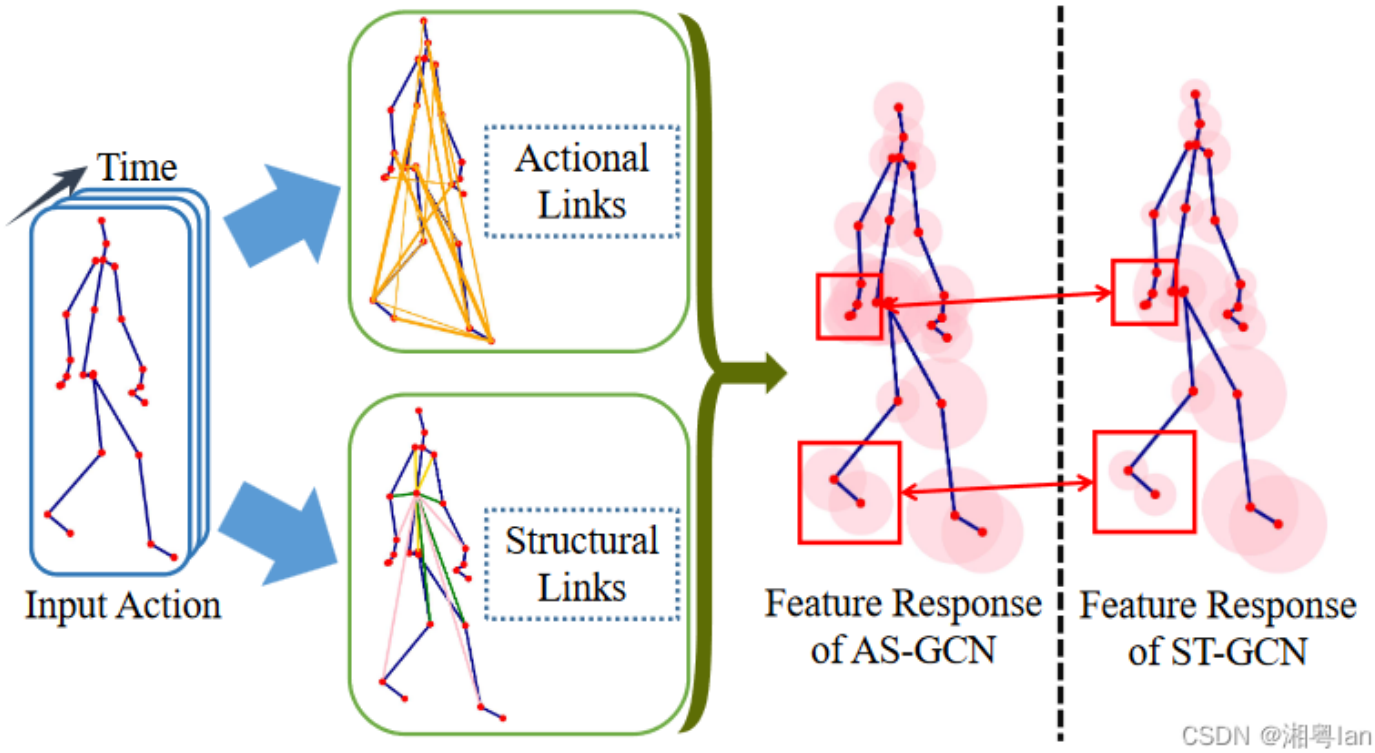
将图卷积应用于基于人体关键点的动作识别的开山之作





AS-GCN

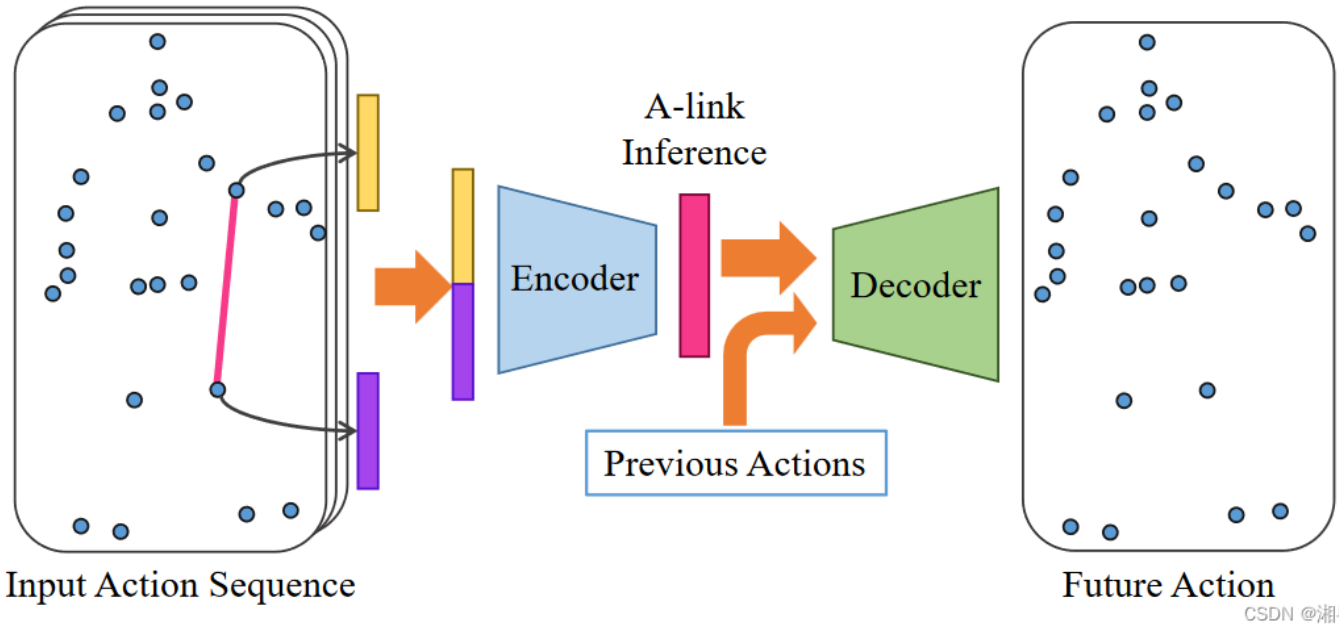
广义边+未来姿势预测：人类的许多动作都需要相隔很远的关节才能协同移动，这导致了关节之间的非物理依赖性。为了捕捉各种动作的对应依赖关系，论文引入了动作链接(A-links)，它由动作激活，可能存在于任意一对关节之间。

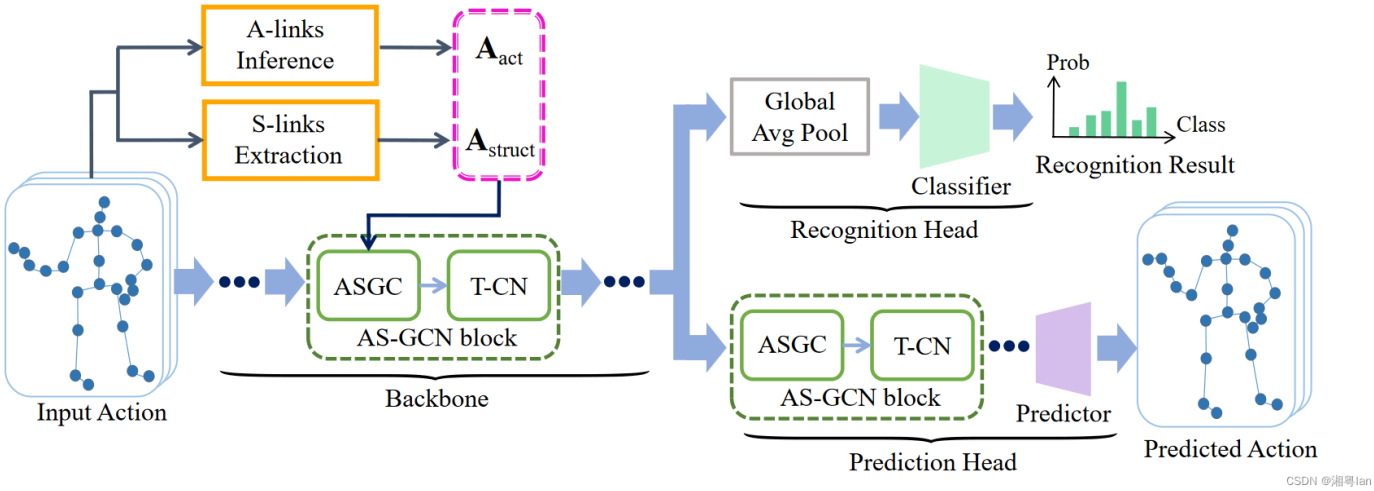


为了从动作中自动推断出A-link，论文开发了一个可训练的A-link推理模块(AIM)，该模块由一个编码器(Encoder)和一个解码器(Decoder)组成：

Encoder：会根据输入信息在训练过程学习任意2 个节点之间的连结，并生成与行为直接相关的A-link。

Decoder：以Encoder 生成的特征(A-link) 推论下个时间点的节点位置。





备注

凡是列出的论文，都是衡量准确度和速度之后的选择：在准确度相比于SOTA没有大幅下降的情况下，更愿意选择速度能超过、达到或接近实时性要求（FPS≥25）的方案。

术语理解参考资料链接

attention→self-attention→transformer

链接一（上半部分）

https://blog.csdn.net/Tink1995/article/details/105012972

链接二（下半部分）

https://blog.csdn.net/Tink1995/article/details/105080033?ops_request_misc=%257B%2522request%255Fid%2522%253A%252216741127041678242861992020%2522%252C%2522scm%2522%253A%252220140713.130102334..%2522%257D&request_id=167411270416782428619920&biz_id=0&utm_medium=distribute.pc_search_result.none-task-blog-2~all~top_positive~default-1-105080033-null-null.142^v71^control,201^v4^add_ask&utm_term=Transformer&spm=1018.2226.3001.4187

空洞卷积（又叫 膨胀卷积）

https://blog.csdn.net/qq_44177768/article/details/123988733

可变形卷积DCN

https://blog.csdn.net/fenglepeng/article/details/121097088?ops_request_misc=%257B%2522request%255Fid%2522%253A%252216741274801678242748887474%2522%252C%2522scm%2522%253A%252220140713.130102334..%2522%257D&request_id=167412748016782427488874&biz_id=0&utm_medium=distribute.pc_search_result.none-task-blog-2~all~top_click~default-1-121097088-null-null.142^v71^control,201^v4^add_ask&utm_term=DCN&spm=1018.2226.3001.4187

全卷积FCN（又叫 反卷积 转置卷积）

https://blog.csdn.net/qq_41760767/article/details/97521397?ops_request_misc=%257B%2522request%255Fid%2522%253A%252216751646041678242979331010%2522%252C%2522scm%2522%253A%252220140713.130102334..%2522%257D&request_id=167516460416782429793310&biz_id=0&utm_medium=distribute.pc_search_result.none-task-blog-2~all~top_positive~default-1-97521397-null-null.142^v72^insert_down3,201^v4^add_ask&utm_term=%E5%85%A8%E5%8D%B7%E7%A7%AF&spm=1018.2226.3001.4187

三种上采样的方式

https://blog.csdn.net/qq_36571422/article/details/122352493?ops_request_misc=%257B%2522request%255Fid%2522%253A%252216751647451680018276778686%2522%252C%2522scm%2522%253A%252220140713.130102334..%2522%257D&request_id=167516474516800182767786&biz_id=0&utm_medium=distribute.pc_search_result.none-task-blog-2~all~sobaiduend~default-1-122352493-null-

null.142^v72^insert_down3,201^v4^add_ask&utm_term=%E4%B8%8A%E9%87%87%E6%A0%B7%E7%9A%84%E5%87%A0%E7%A7%8D%E6%96%B9%E6%B3%95&spm=1018.2226.3001.4187

似然性

https://blog.csdn.net/weixin_45798684/article/details/105841357?ops_request_misc=%257B%2522request%255Fid%2522%253A%252216748200901680022558711 2%2522%252C%2522scm%2522%253A%252220140713.130102334..%2522%257D&request_id=167482009016800225587112&biz_id=0&utm_medium=distribute.pc_search_result.none-task-blog-2~all~sobaiduend~default-2-105841357-null-null.142^v71^control,201^v4^add_ask&utm_term=%E4%BC%BC%E7%84%B6%E6%80%A7&spm=1018.2226.3001.4187

FPN与特征金字塔

https://blog.csdn.net/weixin_55073640/article/details/122627966?ops_request_misc=%257B%2522request%255Fid%2522%253A%252216748031051680018853859 0%2522%252C%2522scm%2522%253A%252220140713.130102334..%2522%257D&request_id=167480310516800188538590&biz_id=0&utm_medium=distribute.pc_search_result.none-task-blog-2~all~top_positive~default-1-122627966-null-null.142^v71^control,201^v4^add_ask&utm_term=FPN&spm=1018.2226.3001.4187

BiFPN

https://blog.csdn.net/qq_38253797/article/details/118439965?ops_request_misc=%257B%2522request%255Fid%2522%253A%252216748231791680018272272 6%2522%252C%2522scm%2522%253A%252220140713.130102334..%2522%257D&request_id=167482317916800182722726&biz_id=0&utm_medium=distribute.pc_search_result.none-task-blog-2~all~top_click~default-2-118439965-null-null.142^v71^control,201^v4^add_ask&utm_term=BiFPN&spm=1018.2226.3001.4187

Focal loss

https://blog.csdn.net/qq_41204464/article/details/122671175?ops_request_misc=%257B%2522request%255Fid%2522%253A%252216748046261678242972631 1%2522%252C%2522scm%2522%253A%252220140713.130102334..%2522%257D&request_id

=167480462616782429726311&biz_id=0&utm_medium=distribute.pc_search_result.none-task-blog-2~all~top_click~default-2-122671175-null-null.142^v71^control,201^v4^add_ask&utm_term=focal%20loss%E6%8D%9F%E5%A4%B1%E5%87%BD%E6%95%B0&spm=1018.2226.3001.4187

Social-STGCNN

https://blog.csdn.net/m0_57541899/article/details/125750780?ops_request_misc=%257B%2522request%255Fid%2522%253A%252216748835331680021303998 9%2522%252C%2522scm%2522%253A%252220140713.130102334.pc%255Fall.%2522%257D&request_id=167488353316800213039989&biz_id=0&utm_medium=distribute.pc_search_result.none-task-blog-2~all~first_rank_ecpm_v1~rank_v31_ecpm-12-125750780-null-null.142^v71^control,201^v4^add_ask&utm_term=%E8%A1%8C%E4%BA%BA%E8%BD%A8%E8%BF%B9%E9%A2%84%E6%B5%8B&spm=1018.2226.3001.4187

其他

行人轨迹和姿态联合预测？人体姿态中也可能包含将来轨迹的运动信息。

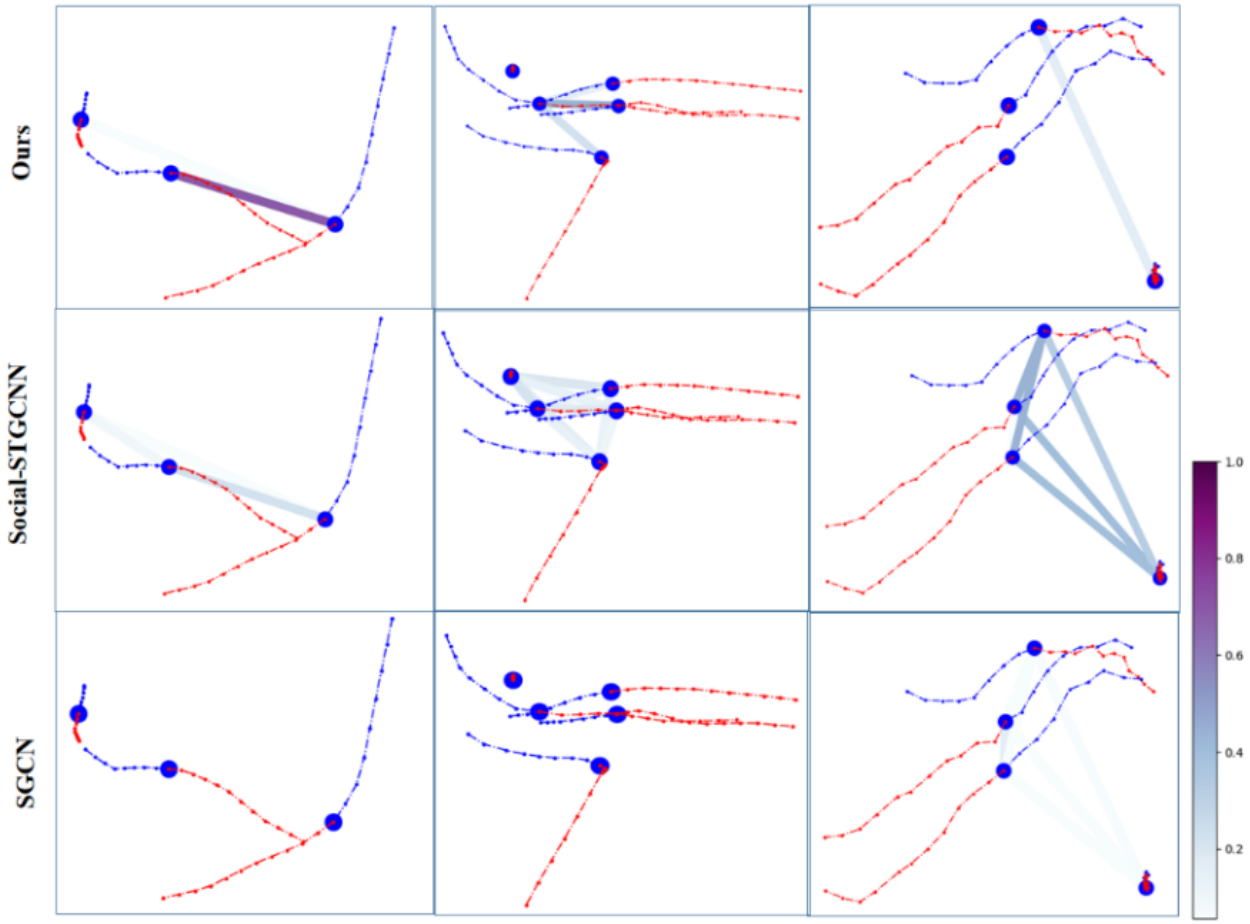


Fig. 7. Visualization of graph relationships in different methods. Each blue node corresponds to one pedestrian and the edge connected to the node itself is not drawn. The blue dotted line represents the historical trajectory (8 frames) of the pedestrian, and the red dotted line represents the ground truth (12 frames). The color of the edge is set according to the colorbar on the right, which describes the weight of the edge. The greater the weight of the edge, the greater the influence between the pedestrians.