计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200700161

基于关键点检测的无锚框轻量级目标检测算法

龚浩田　张　萌
东南大学国家 ASIC 工程中心　南京 210096
(220184705@seu.edu.cn)

摘　要　针对基于关键点的目标检测参数量大、检测框误匹配的问题,提出一种轻量级的基于关键点检测的无锚框目标检测算法。首先将输入图片输入优化过的特征提取算法,通过级联角池化与中心池化,输出 3 个关键点的热力图与它们的嵌入向量;然后通过嵌入向量匹配热力图并画出检测框。文中的创新点在于将 SqueezeNet 中的轻量级模块 firemodule 适配至 CenterNet,并用深度可分离卷积代替主干网的常规卷积,同时,针对 CenterNet 的检测框误匹配问题优化了算法输出形式与训练时的损失函数。实验结果表明,改良后的算法使得原有的 CenterNet 算法模型尺寸缩小为原来的 1/7,同时检测精度与速度较 YOLOv3,CornerNet-Lite 等相同量级的算法仍有所提高。

关键词:目标检测;关键点;无锚框;轻量级;卷积网络
中图法分类号　TP391

Lightweight Anchor-free Object Detection Algorithm Based on Keypoint Detection

GONG Hao-tian and ZHANG Meng
National ASIC Engineering Center,Southeast University,Nanjing 210096,China

Abstract　According to the large number of parameters of key-point object detection network and the problem of mismatching of bounding box,this paper proposes a lightweight key point anchor-free object detection algorithm. It inputs the image into the improved hourglass network to extract features,through the cascade corner pooling module and center pooling module,outputs three key points heatmap and their embedding vectors. At last,it matchs the key points by embedding vectors and draw the bounding box. The innovation of this paper is to applying the firemodule of SqueezeNet in the CenterNet object detection network,and replace the conventional convolution in the backbone with the depth separable convolution. At the same time,aiming at the mismatching bounding box problem in CenterNet,this algorithm adjusts the network's output and loss function. Experiment results show that the model size is reduced to 1/7 of CenterNet,while the accuracy and inference speed are still higher than the same size target detection algorithm like YOLOv3 and CornerNet-Lite.

Keywords　Object detection,Key point,Anchor-free,Lightweight,Convolution network

1　引言

目标检测一直以来都是计算机视觉领域中非常重要的研究方向,主要分为一阶段(one-stage)目标检测和二阶段(two-stage)目标检测。一阶段目标检测是通过回归的方式来输出目标的位置坐标与宽高,代表算法有 YOLO 系列[1-2],CornerNet[3],CenterNet[4]等;二阶段目标检测是先通过区域推荐网络(Region Proposal Network,RPN)来判断前景背景,之后再检测目标的具体位置信息,代表算法有 RCNN[5-7]系列、R-FCN[8]等。

在深度学习目标检测算法中,特征提取(feature extraction)[9]是很重要的一个部分,它是由卷积神经网络(Convolutional Neural Network,CNN)组成的主干网(backbone),主干网的复杂度在很大程度上决定了目标检测算法的参数量与推理时间。更大的网络模型可以提取更丰富的语义信息,达到更好的检测效果,但是也会带来巨大的参数量与计算量,对计算机硬件提出了较高的要求,因此其难以在工业应用领域广泛部署。

2　研究现状

基于关键点检测的无锚框目标检测算法,如 CornerNet,CenterNet 和 ExtremeNet,在一阶段目标检测的基础上,舍弃了 YOLO 系列使用的锚框法,采用回归关键点的方式来输出检测目标的位置与类别,以避免使用锚框时出现参数量大、计算量多的问题,进而大大加快算法速度。但是由于其舍弃了锚框机制,回归出的坐标没有锚框作监督,效果较差,因此在

到稿日期:2020-07-26　返修日期:2020-12-01　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
通信作者:张萌(zmeng@seu.edu.cn)

特征提取时若要提取到丰富的语义信息，就需要融合原图的位置信息来保证检测的准确性[10-11]。

基于关键点的目标检测算法几乎都采用了堆叠沙漏网络(hourglass)[12]，通过不断卷积提取深层语义信息，再通过上采样并融合卷积前相同尺度的特征图来防止浅层信息丢失，从而提取到深层浅层相结合的丰富语义信息[13-14]。但是这些算法的问题是网络较大，参数较多，训练出的目标检测模型近 800 MB，给硬件部署带来很大的挑战。

为了减小算法模型，提高算法推理速度，轻量级的卷积神经网络层出不穷，经典的算法有 MobileNet[15]，SqueezeNet[16]，GoogLeNet[11]，DenseNet[17]等。在不增加模型大小和计算成本的情况下，研究者构建了更高效的算法结构[18]，如深度可分离卷积(depthwise separable convolution)体现了因式分解的思想，将标准卷积分解为深度卷积(depthwise convolution)，再用 1 * 1 卷积核进行常规卷积，这样构建的网络参数较少，可以大大降低算法的复杂度并提高推理速度。

3 相关工作

CenterNet 算法是目前基于关键点的目标检测算法中性能最好的算法，但是由于主干网的大量参数，使得模型大小达到了 800 MB 以上，因此本文在 CenterNet 算法的基础上提出了一种改进的高性能轻量级的无锚框目标检测算法。其轻量级实现原理是使用 SqueezeNet 中的 firemodule 代替 hourglass 中的残差模块，用深度可分离卷积代替 hourglass 中的常规卷积，从而达到减少参数和计算量的目的。同时结合 CenterNet 的思路，以两个角点加目标中心点的形式检测物体，从而使用更少的参数，提高了检测算法的推理速度，压缩了检测算法的模型大小。

同时我们在实验中发现，CenterNet 会出现大框将小框包围的错误匹配框。如图 1 所示，A1，A2，A3 均为同类物体，当同类物体规则排列时，会出现中间物体的中心点恰好在左上物体的左上角点与右下物体的右下角点的中心位置，造成图中黑框的误匹配情况，因此本算法在关键点检测的中心点上增加了嵌入向量进行匹配，以调整算法输出，使得算法针对同类物体也会对中心点进行区分，进而约束最终的回归框来避免误匹配的情况。

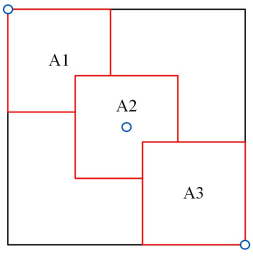


图 1 检测框误匹配

Fig. 1 Mismatching bounding box

4 网络模型设计

为了解决以堆叠沙漏网络为主干网的无锚框目标检测算法参数较多、模型过大，同时推理时间过长的问题，本文根据 CenterNet 提出了 CenterNet_Squeeze_center 检测算法，在主干网中应用了深度可分离卷积并适配 firemodule 模块，同时优化了算法输出形式。

4.1 主干网优化

将 firemodule 适配进 hourglass，其中 firemodule 的结构如图 2 所示，首先用 1 * 1 卷积缩减输入特征图通道数，然后用 1 * 1 卷积与 3 * 3 卷积进行提取特征，再与原特征图进行融合，实现浅层深层的语义信息融合。用深度可分离卷积代替 hourglass 中的常规卷积，从而得到轻量级的 hourglass 特征提取网络。hourglass 特征提取网络的结构如图 3 所示，不同大小的框表示不同的特征图，首先将输入图片逐步卷积提取特征，再将特征图重复上采样，同时为了增加原始特征，防止信息丢失，采用了 ResNet[18-19]的思路，将未卷积的特征图也连接到上采样后的同样尺寸的特征图上以进行通道上的叠加。

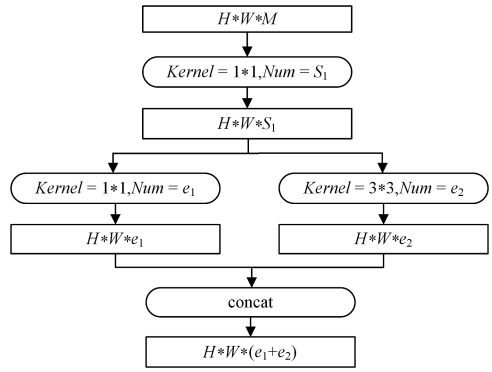


图 2 Firemodule 的结构

Fig. 2 Structure of firemodule

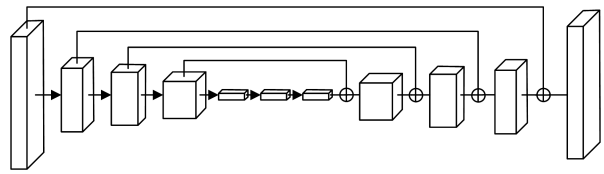


图 3 沙漏网络结构

Fig. 3 Structure of hourglass network

4.2 级联角池化与中心池化

对于经过主干网提取特征后得到的特征图，将其输入至级联角池化模块用于预测左上、右下两个角点。以左上角点为例，先在特征图 x 方向从右向左依次取最大值，第一个最大值处即为物体的上边界；然后从此处向物体内部(y 方向向下)取特征响应值最大值，并与边界点的响应值相加，再对 y 方向从下向上依次取最大值，第一个最大值处即为物体的左边界；再由此向物体内部(x 方向向右)取特征响应值最大值，并与边界点的响应值相加，由得到的这两个融合了内部信息的边界响应点即可定位物体的左上角点。同理，可以得到物体的右下角点。

对于经过主干网提取特征后得到的特征图，将其输入至中心池化模块用于预测物体的中心点。首先在 x 方向先从右

向左依次取最大值，再从左向右依次取最大值，得到特征图1；然后在原特征图的 y 方向先从下向上依次取最大值，再从上向下依次取最大值，得到特征图2；将特征图1与特征图2相加，便可确定目标的特征响应最大值的位置，即目标中心点。

4.3 网络检测原理

算法整体检测结构如图4所示，以优化后的轻量级 hourglass 作为主干网，对输入图片进行特征提取，输入尺寸为 511 * 511，输出特征图(feature map)的尺寸为 128 * 128，然后将特征图输入至两个模块，分别为级联角池化模块和中心池化模块，如图5所示。

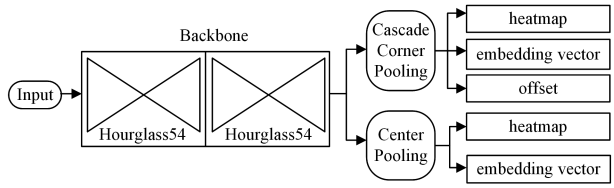


图4 整体检测结构

Fig. 4 Structure of detection network

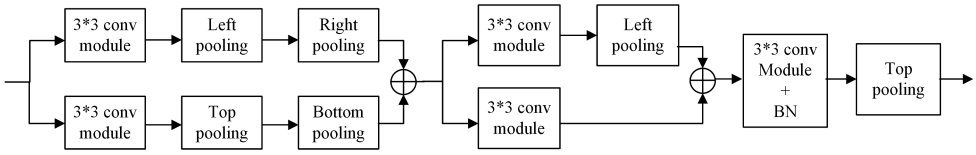


图5 中心池化模块的结构

Fig. 5 Structure of center pooling module

对得到的特征图通过角点预测模块，得到算法输出：左上和右下的角点热力图(heatmap)、嵌入向量(embedding vector)、偏移量(offsets)。角点热力图的尺寸为 128 * 128 * C，此处 C 为通道数，其数值等同于数据集中的目标类别，以 MS COCO 数据集为例，$C=80$。嵌入向量的尺寸为 128 * 128 * 1，相当于每组左上、右下角点的标签，用于最终的匹配角点；偏移量的尺寸为 128 * 128 * 2，因为轻量级 hourglass 有降采样和上采样的过程，取整操作对于像素会有一定程度的偏移，偏移量用于在 x 与 y 方向修正这一误差，左右类别共享一组偏移量。

将算法输出的热力图、嵌入向量、偏移量结合中心点热力图与嵌入向量进行输出解码，将左上角点、右下角点以嵌入向量进行配对，角点的嵌入向量之差小于阈值保留框，然后以22500像素为面积的分界，将小框平均划分为9块，大框平均划分为25块，在划分的中间格检测是否含有中心点热力图，若存在中心点热力图，则保留此框作为最终输出的预测框(bounding box)。

5 损失函数优化

本文损失函数在 CenterNet 损失函数的基础上进行了优化，具体如式(1)所示，它由3部分组成，其中 $\alpha=\beta=0.1$，$\gamma=1$，分别代表检测损失、推拉损失与偏移量损失。其中检测损失的公式如式(2)和式(3)所示，其使用预测到的关键点与标签中的关键点共同计算损失，其中 $\alpha=2$，$\beta=4$，p_{ci} 表示输出热力图中坐标 (i,j) 处为类别 C 的概率，$y_{c_{ij}}$ 表示标签的热力图。

$$L=L_{\text{det}}+\alpha L_{\text{pull}}+\beta L_{\text{push}}+\gamma L_{\text{offset}} \tag{1}$$

$$L_{\text{det}}=-\frac{1}{N}\sum_{c=0}^{C}\sum_{i=1}^{H}\sum_{j=0}^{W}(1-p_{c_{i,j}})^{\alpha}\log(p_{c_{i,j}}),\text{当 } y_{c_{i,j}}=1 \tag{2}$$

$$L_{\text{det}}=-\frac{1}{N}\sum_{c=0}^{C}\sum_{i=1}^{H}\sum_{j=0}^{W}(1-p_{cij})^{\beta}p_{c_{i,j}}^{\alpha}\log(1-p_{c_{i,j}}),\text{当 } y_{c_{i,j}}\neq 1 \tag{3}$$

推拉损失的公式如式(4)和式(5)所示，pull 损失用于使预测到的同个物体的左上、右下、中心3个关键点的嵌入向量尽量相等；push 损失用于使不同物体的关键点的嵌入向量差值尽量大，其中 $\boldsymbol{e}_{t_k}$ 为左上角的嵌入向量，$\boldsymbol{e}_{b_k}$ 为右下角的嵌入向量，$\boldsymbol{e}_{c_k}$ 为中心点的嵌入向量，$\boldsymbol{e}_k$ 为左上、右下、中心点嵌入向量的均值。

$$L_{\text{pull}}=\frac{1}{N}\sum_{k=1}^{N}[(\boldsymbol{e}_{t_k}-\boldsymbol{e}_k)^2+(\boldsymbol{e}_{b_k}-\boldsymbol{e}_k)^2+(\boldsymbol{e}_{c_k}-\boldsymbol{e}_k)^2] \tag{4}$$

$$L_{\text{push}}=\frac{1}{N(N-1)}\sum_{k=1}^{N}\sum_{j=1,j\neq k}^{N}\max(0,1-|\boldsymbol{e}_k-\boldsymbol{e}_j|) \tag{5}$$

偏移量损失的公式如式(6)和式(7)所示，由于卷积会造成输出的特征图分辨率比原始分辨率低，假设其下采样率为 n，原图上 (x,y) 点映射到特征图上的位置就为 $(\lfloor x/n\rfloor,\lfloor y/n\rfloor)$，当在特征图上预测出一个角点后，映射回原图时就会有所误差，即式(6)所示的偏移量，虽然可能只有几个像素的偏差，但是在小目标的情况下，这会严重影响 IoU。$\hat{O}_k$ 为预测坐标偏差，训练使用的损失函数为 SmoothL1。

$$O_k=(\frac{x_k}{n}-\lfloor\frac{x_k}{n}\rfloor,\frac{y_k}{n}-\lfloor\frac{y_k}{n}\rfloor) \tag{6}$$

$$L_{\text{offset}}=\frac{1}{N}\sum_{k=1}^{N}\text{SmoothL1}Loss(O_k,\hat{O}_k) \tag{7}$$

6 实验结果及分析

6.1 实验环境

本文实验使用主流深度学习框架 Pytorch0.4 搭建神经网络，使用一个 Tesla V100 GPU 显卡进行模型训练(文献[4]中使用10块 GPU 进行训练，可以设置更大的 *batch size*，因此可以达到更好的收敛效果并得到更精确的检测框)，数据集使用 MS COCO 数据集，使用 train2014(80 000 幅图像)和 val2014(35 000 幅图像)进行训练。数据集包含80类常见目标类别与多达150万个物体实例，标注为 json 格式文件。本文使用平均精度(average precision)与平均召回率(average recall)作为算法的评价指标，对于不同尺寸的目标通过交并比(Intersection over Union，IoU)计算精度与召回率。

6.2 实验步骤及细节

本文实验是以 CornerNet-Squeeze 为 Baseline，将所提算法与 CornerNet-Squeeze，CenterNet，YOLO-v3 在同等训练条

件下作对比，下面具体介绍本文算法的训练方法。

(1)网络不使用预训练模型，使用随机初始化权重方法开始训练。

(2)对数据集使用标准数据增强，即翻转、缩放、随机裁剪和颜色抖动。

(3)损失梯度下降优化使用 Adam 方法。

(4)由于硬件环境限制，*batch size* 设置为 12，此参数随着 GPU 内存的增大而增大，且参数越大，模型越容易收敛；学习率(learning rate)设置为 0.000 25。

(5)本文在训练过程中使用了一个技巧，即使用了总共 250 000 次迭代代替了 CenterNet 的 450 000 次迭代，同时在迭代 240 000 次时衰减学习率为初始的 1/10，我们发现这种方法可以得到同样的训练结果。

6.3 实验结果的对比

表 1 列出了常见的关键点目标检测算法以及它们在原文中的精度。然而大部分研究在训练模型时使用的 GPU 数量为 8～12，这意味着可以在训练时使用更大的 *batch size*，如 CenterNet 使用了 8 块 Tesla V100 GPU，使得 *batch size* 可以设置为 128，更大的 *batch size* 意味着每批次训练可以根据更多数据进行拟合，从而获得拟合效果更好的算法模型。由于硬件设备的限制，在训练中我们只能使用一块 Telsa V100 GPU，*batch size* 只能设置为 12，所以我们同时训练了表 1 中的其他模型以作对比，所有模型均在相同环境、相同训练条件下进行。表 2 列出了本文提出的算法与其他算法的性能。

表 1 常见的检测器的性能

Table 1 Performance of common detector

Algorithm	Mean average precion/%	Inference Speed/ms	Model size/MB	Parameter/M
YOLOv3	33	49	246	61.5
CenterNet	44.9	224	802	210
CornerNet_ Squeeze[20]	34.4	30	122	31

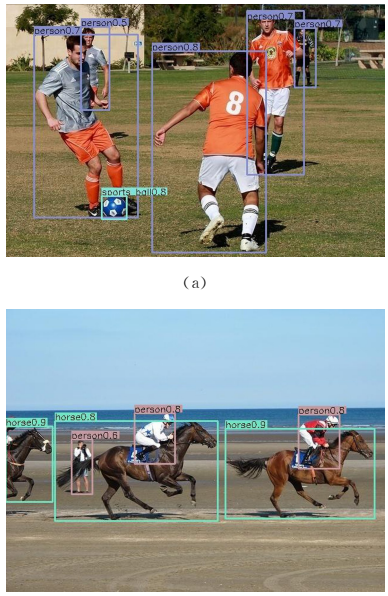
由于本文算法是在 CenterNet 的基础上进行轻量化，精度较 CenterNet 会有一定的下降，这是无法避免的，但是其推理速度得到很大提升，参数量大幅减少。与同量级的检测算法 CornerNet-Squeeze 相比，CenterNet-Squeeze 由于增加了目标中心点作为标定，使得检测精度在不降低推理速度的前提下依然有所提高。在此基础上，本文创新地在中心点增加了嵌入向量来避免误匹配问题，使得精度得到进一步提高，即表 2 中的 CenterNet_Sq_ct。

表 2 各算法的实验结果对比

Table 2 Comparison of experimental results of each algorithm

Algorithm	Mean average precion/%	Inference Speed/ms	Model size/MB	Parameter/M
YOLOv3	30.5	52	246	61.5
CenterNet	40.1	276	802	210
CornerNet_ Squeeze	31.2	39	122	31
CenterNet_ Squeeze	32.5	39	129	32
CenterNet_ Sq_ct	33.1	40	133	34

由表 2 可以看出，本文提出的 CenterNet_Squeeze 轻量级检测算法是基于 CenterNet 算法进行的轻量化改进，使得算法模型减小为原来的 1/5，参数量减少为原来的 1/7，推理速度提高了 7 倍，虽然精度有一定程度的下降，但是相比相同规模的轻量级算法 CornerNet_Squeeze，其精度依然提高了 1.3%；同时在本文训练条件下的对比实验中，CenterNet_Sq_ct 的精度与推理速度均优于 YOLOv3。检测结果示例如图 6 所示。



(a)

(b)

图 6 检测结果示例

Fig. 6 Sample diagrams of experimental results

结束语 本文提出了轻量级目标检测算法 CenterNet_Sq_ct，使得算法模型减小为原来的 1/5，参数量减少为原来的 1/7，推理速度提高了 7 倍，与同等量级的检测算法相比，本算法的精度与推理速度均优于 YOLOv3。本文在 MS COCO 数据集上验证了所提算法的良好效果。

由于本文中的实验环境无法提供足够的 GPU 进行训练，算法无法拟合到最佳状态；同时在测试过程中发现，在遇到多类相似目标密集的复杂场景时，检测效果欠佳，这也是关键点算法检测的通病。未来的研究方向是通过对关键点进行后处理，来避免同类关键点聚成一片的情况，从而得到更好的检测效果。

参 考 文 献

[1] REDMON J, FARHADI A. Yolov3: An incremental improvement[J]. arXiv: 1804.02767.

[2] REDMON J, FARHADI A. YOLO9000: Better, Faster, Stronger [C]// IEEE Conference on Computer Vision & Pattern Recognition. 2017: 6517-6525.

[3] LAW H, DENG J. Cornernet: Detecting objects as paired keypoints[C]// Proceedings of the European Conference on Computer Vision. 2018: 734-750.

[4] DUAN K W, BAI S, XIE L X. CenterNet: Keypoint Triplets for

Object Detection[C]// Proceedings of the European Conference on Computer Vision. 2019.
[5] REN S Q, HE K M, GIRSHICK R. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6): 1137-1149.
[6] GIRSHICK R. Fast R-CNN[C]// Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). 2015: 1440-1448.
[7] HE K, GKIOXARI G, DOLLAR P, et al. Mask r-cnn[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 2961-2969.
[8] ZHOU X, WEI G, FU W L, et al. Application of deep learning in object detection[C]// 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS). IEEE, 2017.
[9] MURPHY K P. Object detection and localization using local and global features[J]. Toward Category Level Object Recognition, 2006, 12(1): 382-400.
[10] WANG W, SHEN J, SHAO L. Video Salient Object Detection via Fully Convolutional Networks[J]. IEEE Transactions on Image Processing, 2017, 27(1): 38-49.
[11] YANG J, LIU Q S, ZHANG K H. Stacked Hourglass Network for Robust Facial Landmark Localisation[C]// IEEE Conference on Computer Vision & Pattern Recognition Workshops. IEEE Computer Society, 2017.
[12] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 580-587.
[13] XIAO B, WU H, WEI Y. Simple baselines for human pose estimation and tracking[C]// ECCV. 2018: 472-478.
[14] SANDLER M, HOWARD A, ZHU M L, et al. MobileNetV2: Inverted Residuals and Linear Bottlenecks [C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 4510-4520.
[15] IANDOLA F N, HAN S, MOSKEWICZ M W, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and < 0.5 MB model size[J]. arXiv, 1602.07360.
[16] SZEGEDY C, LIU W, JIA Y, et al. Going Deeper with Convolutions[J]. arXiv: 1409.4842.
[17] HUANG G, LIU Z, LAURENS V D M, et al. Densely Connected Convolutional Networks[J]. arXiv: 1608.06993, 2016.
[18] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// CVPR. 2016.
[19] HE K, ZHANG X, REN S, et al. Identity Mappings in Deep Residual Networks[C]// European Conference on Computer Vision. Cham: Springer, 2016.
[20] LAW H, TENG Y, RUSSAKOVSKY O, et al. CornerNet-Lite: Efficient Keypoint Based Object Detection [J]. arXiv: 1904.08900.

GONG Hao-tian, born in 1996, postgraduate. His main research interests include deep learning and computer vision.

ZHANG Meng, born in 1964, Ph.D, associate professor, Ph.D supervisor. His main research interests include deep learning, machine learning, digital signal processing, digital communication systems, wireless sensor networks, digital integrated circuit design, information security and assurance, etc.