文献标识码：B 文章编号：1003-0492（2022）08-070-06 中图分类号：TP274

基于任务关联的端到端目标检测算法

Task Aligned End-to-End Object Detection

★王　宏（北京轩宇空间科技有限公司，北京 100086）
★曾　峥（中国空间技术研究院北京控制工程研究所，北京 100086）
★徐奕男，谷晓琳（北京轩宇空间科技有限公司，北京 100086）

摘要：对预测结果准确排序是端到端目标检测的关键。已有的端到端检测器将分类和定位当作独立任务，减少了两者之间的关联，导致利用分类分数排序的结果不可靠，降低了检测性能。针对上述问题，本文从样本选择、损失函数和网络结构三个方面进行了优化，提高了两者之间的一致性。首先利用分类和定位的排序结果计算样本选择的代价矩阵，并优先考虑分类和定位一致性大的样本作为正样本。另外，使用基于任务关联的损失函数训练分类器，学习同时表示目标分类精度和定位准确度的分数。考虑到分类和定位对特征需求的差异，在头部检测网络中引入特征对齐层，缓解了分类和定位之间的冲突。在COCO数据集上，基于任务关联的端到端目标检测算法的性能优于许多优秀的检测器。

关键词：目标检测；样本选择；损失函数；相关性

Abstract: Accurate ranking of prediction results is the key to end-to-end object detection. Existing end-to-end detectors treat classification and localization as independent tasks, reducing the correlation between them, and resulting the unreliable results ranking by classification scores, which reduces the detection performance. To solve the above problem, we try to improve the correlation between classification and localization form label assignment, loss function and network structure. First, the cost matrix of sample selection is calculated by using the sorting results of classification and location, and the samples with high consistency of classification and location are given priority as positive samples. We propose a task-aligned loss function to train the classifier, which aims to learn a score that can simultaneously represent the accuracy of object classification and localization. We introduce a feature alignment layer in the head detection network, which alleviates the conflict between classification and localization at the feature extraction level. On COCO datasets, the end-to-end object detection algorithm proposed in this paper outperforms many excellent detectors.

Key words: Object detection; Label assignment; Loss function; Correlation

1　引言

目标检测是计算机视觉的一项基础研究，对图像理解、信息提取等有着重要意义，广泛应用于自动驾驶、智能监控等多个领域[1~2]。近年来，基于深度学习的目标检测算法[3~9]取得了很大进步，这些检测器在推理阶段预测密集的检测结果，需要使用非极大值抑制算法[10]对冗余的检测结果进行筛选，不能实现完全端到端的目标检测。

DETR[11]利用Transformer解码结构替换了基于卷积的头部网络，对特征金字塔提取的特征进行解码，输出分类分数和边界框，并利用分类和定位的损失和作为代价函数，为每个目标选择一个正样本训练网络。DETR在训练和推理阶段均无需使用非极大值抑制算法，实现了完全端到端的目标检测。DeFCN[12]基于全卷积实现了完全端到端的目标检测。DeFCN提出了基于预测信息的样本选择策略，利用匈牙利算法为每个目标选择一个合适的样本。DeFCN设计了一种3D-Max滤波器，利用多尺度特征提高了卷积在局部区域的可分辨性。

端到端目标检测器利用分类分数对检测结果排序，选择前N个分数最高的输出。由于目标检测是多任务机制，不仅需要对目标进行分类，还需要回归准确的边界框。已有的端到端检测算法将分类和定位当作两个独立的任务进行训练，减少了分类和定位之间的关联性，导致分类分数高的地方往往不是定位最准确的，如图1所示。针对上述问题，本文提出了一种基于任务关联的端到端目标检测器，从样本选择、损失函数和网络结构三

个方面进行了改进，提高分类和定位的相关性。

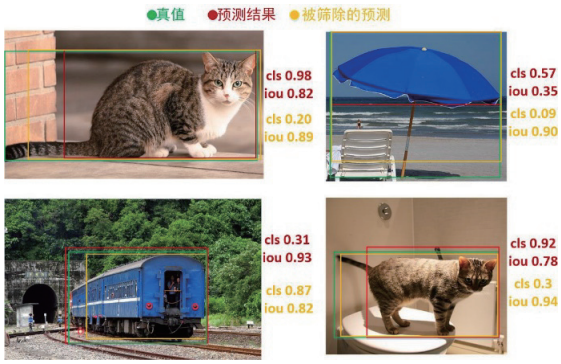

图1 端到端目标检测分类分数和回归结果不一致

大多数目标检测算法只利用位置信息选择样本，如RetinaNet[8]使用基于锚的样本选择方法和FCOS[9]利用基于中心点距离的样本选择方法。基于位置信息的样本选择方法为每个目标选择多个正样本，为网络训练提供了丰富的特征信息。但是由于目标检测是分类和定位联合的任务，只考虑位置信息的样本选择方法和网络优化的目标不一致，导致冗余的高分检测框需要使用非极大值抑制算法对检测结果进行筛选。OneNet[13]指出同时考虑分类和定位信息的一对一样本选择方法是移除非极大值抑制算法，实现端到端目标检测的关键，并提出基于最小代价的样本选择方法，利用分类损失和定位损失构建代价矩阵，为每个目标选择代价最小的样本作为正样本。DETR和DeFCN在选择样本时也同时考虑了分类和定位信息，并为每个目标选择一个合适的正样本。不同于上述三种基于预测信息的样本选择方法，本文利用分类和定位的排序结果计算样本选择的代价矩阵，并引入分类和定位的相关性计算权重，优先选择一致性高的样本作为正样本。

为了进一步提高分类和定位的关联，本文使用基于任务关联的损失函数训练分类分支，学习可以同时表示目标分类精度和定位准确度的分数。已有的目标检测算法证明了在损失函数中引入分类和回归的相关性，有利于提高目标检测算法的性能。文献[14]提出广义Focal Loss（General Focal Loss，GFL）用于训练基于分类和回归的联合分数，并用该分数对预测结果进行排序。VFNet[15]提出了一种基于位置感知的分数，联合表征目标的分类和回归质量。上述两种方法都是基于多对一的样本选择方法提出的，并不适用于端到端目标检测器。与上述方法不同，本文使用一对一目标选择策略，每个目标只有一个正样本，利用任务关联的损失函数训练分类器，使网络向分类和定位一致性好的方向优化。

考虑到分类和定位对特征的需求存在差异，分类倾向于选择旋转和平移不变的特征，而回归对旋转和平移更加敏感。本文在头部检测网络中添加了特征对齐层，缓解分类和定位在特征层面的冲突，进一步加强了两者的关联性。

2 基于任务关联的端到端检测算法

2.1 基于分类和定位联合的样本选择方法

基于预测信息的样本选择方法[13]，被证明是实现端到端目标检测的关键。基于预测信息的样本选择方法首先根据网络的预测信息计算与目标真值的分类代价和回归代价，并利用两种代价构造损失函数，如式（1）所示：

$$C_{\mathrm{f}} = \lambda_{\mathrm{cls}} C_{\mathrm{cls}} + \lambda_{\mathrm{IoU}} C_{\mathrm{IoU}} \tag{1}$$

其中，λ_{IoU}和λ_{cls}分别为回归损失和分类损失的权重系数；C_{IoU}表示网络预测的回归框与目标框的回归损失，即IoU损失；C_{cls}表示分类损失，由交叉熵计算得到。

基于预测信息的样本选择方法将分类和定位信息当作独立的变量构造代价函数，在分类和定位不一致时，容易产生分歧，如图2所示，同一个目标在分类分数高的地方，定位不准确；但定位准确的地方分类分数较低。

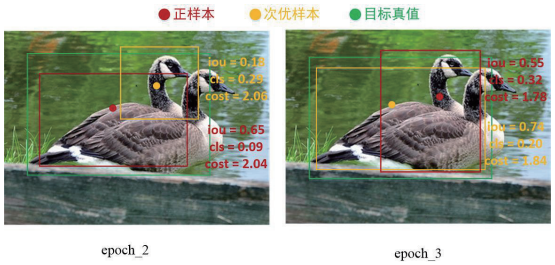

图2 基于预测信息的样本选择方法在不同迭代中选择的样本示意图

为了减缓分类和定位的分歧，本文优先选择两者一致性高的样本作为正样本，如式（2）所示：

$$\hat{\pi} = \underset{\pi \in \Pi_G^N}{\arg\max} \sum_{i}^{G} C_{i,\pi(i)}, \tag{2}$$

其中，$C_{i,\pi(i)}$表示第i个目标的代价，计算过程如式（3）所示：

$$C_{i,\pi(i)} = t_{\pi(i)} \cdot o_{\pi(i)} \tag{3}$$

$C_{i,\pi(i)}$由权重系数$t_{\pi(i)}$和联合质量$o_{\pi(i)}$两部分组成，其中联合质量的计算方法如式（4）所示：

$$o_{\pi(i)} = \exp\left(-\theta \cdot \left(R^{\text{iou}}_{\pi(i)} + R^{\text{cls}}_{\pi(i)}\right)\right) \tag{4}$$

$R^{\text{iou}}_{\pi(i)}$（$R^{\text{iou}}_{\pi(i)} = 1,2,\text{L},n$）表示回归分数从高到低的排序结果，$R^{\text{cls}}_{\pi(i)}$表示分类分数从高到低的排序结果。样本的联合分数越高，证明其一致性越好，被选为正样本的概率越大，θ是调节因子。$\mathbf{t}_{\pi(i)}$为权重系数，根据分类和定位的相关性计算得到，如式（5）所示：

$$t_{\pi(i)} = \exp\left(s^{\gamma}_{\pi(i)}\right), \tag{5}$$

其中，$s_{\pi(i)} = \dfrac{\min\left(\hat{p}_{\pi(i)}, \hat{q}_{\pi(i)}\right)}{\max\left(\hat{p}_{\pi(i)}, \hat{q}_{\pi(i)}\right)}$，$\hat{p}_{\pi(i)}\left(c_i\right)$，表示检测网络预测的第$i$个目标的类别分数。$\hat{q}_{\pi(i)}$表示网络预测的边界框与目标真实边界框的IoU分数。γ是调节因子。分类分数和回归分数越接近，样本的权重越大。本文使用匈牙利算法，根据代价矩阵$C_{i,\pi(i)}$，为每个目标选择一个合适正样本。

2.2 基于任务关联的分类损失函数

大多数目标检测器使用Focal Loss[8]训练分类器，Focal Loss的定义如式（6）所示：

$$\begin{cases} -\alpha(1-\boldsymbol{p})^{\beta} \ln \boldsymbol{p} & \boldsymbol{y}=1 \\ -(1-\alpha)\boldsymbol{p}^{\beta} \ln(1-\boldsymbol{p}) & \boldsymbol{y}=0 \end{cases} \tag{6}$$

其中，$\boldsymbol{y}$是样本类别标签，如果是正样本，则$\boldsymbol{y}=1$；反之$\boldsymbol{y}=0$。Focal Loss只考虑了分类分数对分类器的影响，由于端到端目标检测器使用分类分数对检测结果排序，分类分数不仅表示目标分类的结果，还需要兼顾回归是否准确。本文提出了一种基于任务关联的分类损失函数，利用联合分数替代原来的二值标签，如式（7）所示：

$$L_{cls} = \sum_{i=1}^{N_{pos}} o_i \cdot BCE\left(\hat{p}_i, o_i\right) + \sum_{j=1}^{N_{neg}} \hat{p}_j^{\sigma} BCE\left(\hat{p}_j, 0\right), \tag{7}$$

其中，o_i表示第i个正样本的联合分数，计算方法如式（6）所示。样本的一致性越高，联合分数越大，样本对网络的贡献越大，网络向分类和定位一致性高的方向优化，进一步提高了两者之间的相关性。σ是调节因子，用于调节负样本的权重。

2.3 基于特征对齐的头部检测网络

目标检测中分类任务和回归任务对特征的需求存在差别，分类需要特征具有平移和旋转不变性，能准确地判断目标类别。但定位分支需要准确地回归目标边界框，对平移和旋转敏感。为了在特征层面减少分类和定位任务的冲突，增加两者的相关性，本文在头部检测网络中引入特征对齐网络层，首先利用分类特征学习特征偏移量，通过可形变卷积网络获得对齐后的特征x_{aligned}，如式（8）所示：

$$x_{\text{aligned}} = dconv(x_{\text{reg}}, conv(x_{\text{cls}})) \tag{8}$$

其中，x_{reg}表示回归特征，x_{cls}表示分类特征。对齐特征与原始的回归特征组合，获得新的特征用于目标框的回归。特征对齐网络通过可形变卷积改变了原始回归特征的感受野，为回归目标框提供了更多的信息，一定程度上减缓了分类和定位在特征层面的冲突。使用分类特征学习偏移量，并作用于回归特征。特征对齐层作为桥梁，也增加了特征提取层面分类和定位的相关性。具体的网络结构如图3所示。

3 实验分析

3.1 数据集及实验设计

COCO数据集[16]是目标检测领域通用的数据集之一，共80种类别，包含了丰富的自然图像和生活中常见的目标图像，背景复杂且目标数量多，实验充满了挑战性。实验环境的硬件配置为Intel Xeon E5处理器，4块Nvidia TiTAN RTX GPU，操作系统为Ubuntu18.04。本文基于FCOS[9]框架，利用全卷积神经网络构建了端到端的目标检测器，主要包含骨干网络、特征金字塔网络、头部网络三部分。骨干网络主要使用ResNet50和ResNet101[17]，具体的网络实现基于MMDetection框架[18]。

为了方便与优秀的目标检测方法进行对比，在训练过程中，冻结了骨干网络所有BN层的参数，以及骨干网络第一层卷积的所有参数，利用ImageNet数据集预训练的权重初始化骨干网络[21]。所有实验使用随机梯度下降法[19]（Stochastic Gradient Descent，SGD）进行训练，初始化学习率为0.01。训练阶段，输入图像分辨率设置为短边800个像素，长边不超过1333个像素，保持输入图像的宽高比。每次迭代输入图像数目（batch）为16，平均每个GPU输入图像数目为4。实验最大迭代次数（epoch）设为12。学习率在第8和第11次迭代时衰减10倍，分别为0.001和0.0001。

推理时图像的输入分辨率与训练时相同，得到预测结果后，首先在每个类别中取前100个得分最高的结果，对这些结果进行排序，然后取前100个得分最高的

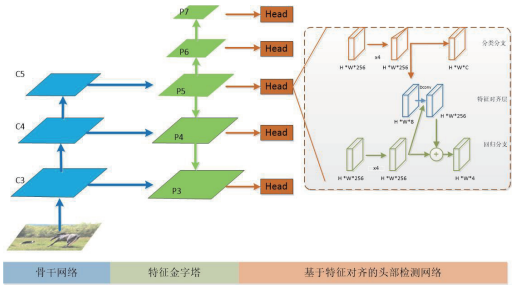

图3 基于联合信息的端到端目标检测网络结构示意图

结果，使用阈值进行筛选，选择大于阈值的结果作为最终输出。

3.2 COCO数据集

3.2.1 超参数的影响

超参数γ和θ是计算权重和联合分数引入的调节因子。本文首先在COCO数据集上通过一组对比实验研究了超参数对检测性能的影响，以COCO中的$mAP_{0.5:0.95}$为评价指标。在推理阶段不使用非极大值抑制算法，直接选择前100个分类分数最高的检测结果输出，得到的结果如表1所示。

表1 超参数对检测性能的影响

	θ=0.01	θ=0.05	θ=0.1	θ=0.15	θ=0.2
γ=0.0	36.1	38.2	33.0	30.2	28.3
γ=3.0	35.9	38.3	32.6	31.1	28.9
γ=5.0	36.4	39.0	33.4	31.5	29.0
γ=8.0	36.0	38.6	32.9	31.0	28.1

θ是计算联合分数的调节因子，当θ=0.01时，检测结果在36%左右。当θ=0.05时，检测性能提升到38%左右。当θ=0.1时，联合分数随着排序结果快速减小，不利于分类器训练，检测结果较差。γ是计算权重系数t时的调节因子，当γ=0.0时，所有样本的权重系数均为1，没有考虑分类和定位的相关性。当γ取值逐渐增大，检测性能得到了提升，γ=5.0时检测性能最高，为39.0%。因此在后续实验中，超参数的取值设置为θ=0.05，γ=5.0。

3.2.2 消融实验

本文通过一组消融实验研究了在样本选择、损失函数和网络结构上的改进对检测性能的影响，如表2所示，“/”之前表示使用NMS的检测结果，“/”后表示不使用NMS的检测结果。实验的基准模型为基于FCOS框架构建的端到端目标检测网络。基准模型使用OneNet提出的基于最小代价的样本选择方法，并使用FcoalLoss训练分类器。如表2第一行所示，基准模型的mAP为38.2%，使用NMS筛选后的结果为39.2%，两者差距为1.0%。当使用本文提出的基于分类和定位信息的联合样本选择方法替换基础端到端检测器的样本选择方法后，AP上升到了38.6%，使用NMS后的AP为39.3%，两者差距为0.7%，缩小了两者的差距，如表2第二行所示。基于任务对齐的分类损失函数替换了Focal Loss后，AP上升到39.0%，经过NMS筛选后的结果为39.4%，两者差距为0.4%，如表2第三行所示。使用基于特征对齐的头部检测网络后，AP提升到了39.8%，而经过NMS筛选后的检测结果为40.1%。可以看出，基于特征对齐的头部检测网络有利于进一步提升端到端目标检测的性能。但从表中可以看出，使用NMS后的均高于未使用NMS的效果，主要因为特征金字塔（FPN）网络在特征提取上存在重复，下一步的工作中，我们将进一步研究网络结构对稀疏输出的影响。

表2 不同的改进模块对检测性能的影响

样本选择策略	联合分数	头部检测网络	mAP (%)	AP50 (%)	AP75 (%)
x	x	x	39.2/38.2	57.7/54.5	42.8/42.1
√	x	x	39.3/38.6	58.3/56.0	42.3/42.6
√	√	x	39.4/39.0	58.1/57.6	43.5/42.7
√	√	√	40.1/39.8	58.6/58.1	43.7/43.1

3.2.3 与其他优秀算法的对比

本实验对比了本文提出算法与其他优秀的目标检测算法的性能，如表3所示。对本文提出的算法的检测结果进行了可视化，如图4所示。对密集目标检测器，我们使用非极大值抑制算法对检测结果进行筛选。

CenterNet[20]基于中心点距离选择与目标中心点最近的样本作为正样本，其余为负样本。在训练过程中，采用高斯权重衰减了目标框内的负样本权重。距离目标框中心点越近的负样本，权重越小。FCOS[9]是经典的一阶段检测器，将中心点落入目标框内的所有候选样本标记为正样本，其余为负样本。另外，FCOS设计了一个独立的分支预测回归质量，用来抑制低回归框对检测性能的影响。VFNet[15]提出了一种基于定位感知的分类分数（IACS），并利用Varifocal Loss作为损失函数进行训练。另外，为了提高回归框的准确度，VFNet在网络结构中增加了回归框的修正分支，用来进一步修正回归的预测值，使其更加准确。对于端到端的目标检测算法，我们选择了前100个类别分数最高预测结果，通过分数阈值T_{score}筛选出高于阈值的结果作为最终输出。OneNet[13]提出了基于最小代价的样本选择策略，为每

个目标选择一个代价最小的样本训练网络。DeFCN[12]是经典的端到端目标检测算法，提出了一种基于预测信息的一对一样本选择策略。另外，DeFCN提出了3D-Max滤波器，利用多尺度特征提高卷积在局部区域的可分辨性。为了对网络模型提供足够的特征信息，DeFCN使用多对一的样本选择策略，选择多个正样本对分类分支进行辅助训练，进一步提高了检测性能。

表3 与其他优秀算法的对比实验

方法	骨干网络	mAP (%)	AP50 (%)	AP75 (%)
CenterNet	ResNet-50	35.0	53.5	37.2
FCOS	ResNet-50	38.7	57.2	41.7
VFNet	ResNet-50	41.6	62.5	48.1
OneNet	ResNet-50	35.6	53.6	38.2
DeFCNPOTO	ResNet-50	38.0	55.2	41.4
DeFCNPOTO+3DMax	ResNet-50	39.8	57.0	42.9
DeFCN	ResNet-50	41.1	59.5	45.6
Ours	ResNet-50	39.8	57.6	42.7

本实验同时对比了密集目标检测器和端到端的目标检测器的检测性能。从表3可以看出，本文提出的算法比CenterNet高出4.8%AP，比FCOS高出1.1%，但落后于VFNet。FCOS和VFNet使用多对一的样本选择策略，相比一对一的采样策略，提供了更加丰富的特征信息用于模型训练，但都需要使用NMS对冗余的检测结果进行筛选。FCOS将落入目标框内的所有样本，都选为正样本，增加了低质量样本对模型的影响。虽然使用一个分支对低质量回归框抑制，但该分支独立进行训练，缺少了与分类和回归的关联，作用较小。VFNet使用自适应阈值的样本选择策略，每个目标选择大于阈值的样本，样本分布较为均匀，另外VFNet增加了对回归框修正的网络分支，对预测结果进行修正。相比FCOS和VFNet,，本文提出的算法没有增加额外的分支对预测结果进行修正，并移除非极大值抑制算法。

本文提出的算法比OneNet高出4.2%AP，与只使用POTO样本选择策略的DeFCN相比，取得了更好的检测性能，但仍然低于完整的DeFCN。从表3中可以看出，DeFCN使用POTO样本选择策略与3D-Max滤波器后的检测效果与本文提出的算法性能相当。当DeFCN增加了多对一样本选择策略对分类器进行辅助训练时，检测性能有了明显提升，多对一的样本选择策略可以为网络训练提供更加丰富的特征信息，基于预测信息的多对一样本选择策略对分类器更加友好，也是本文下一步重点研究的内容之一。

为了进一步证明本文提出方法的鲁棒性和有效性，我们在不同的骨干网络上进行了实验，如表4所示。使用较大的骨干网络ResNet101时，检测性能为42.0%，比使用ResNet50提高了2.2%。当使用可形变卷积（Deformable Convolutions Network，DCN）替代传统卷积后，上升到了44.0%。使用更大的骨干网络可以提供更加丰富的语义信息，因此检测性能也更加优越，通过这组实验，也证明了本文提出的算法可以应用于不同的骨干网络上，具有较好的鲁棒性和有效性。

表4 不同的骨干网络对检测性能的影响

骨干网络	mAP （%）	AP50 （%）	AP75 （%）
ResNet-50	40.0	57.6	42.7
ResNet-101	42.0	60.9	44.8
ResNet-101-DCN	44.0	61.2	47.7

4 结语

端到端目标检测器在推理阶段直接选择前N个分类分数最高的结果作为最终结果输出，移除了非极大值抑制算法。已有的端到端检测算法将分类和定位当作两个独立的任务，减少了分类和定位之间的关联性，导致分类分数高的地方往往不是定位最准确的，造成检测性能的下降。针对上述问题，本文提出了一种基于分类和定位联合的端到端目标检测器，从样本选择、损失函数和网络结构三个方面进行了改进，首先在样本选择的代价函数中引入分类和定位的相关性计算权重，优先选择相关性大的样本为正样本。为了进一步提高分类和定位的一致性，本文利用基于任务对齐的损失函数训练分类器，学习可以同时表征分类准确度和定位精度的联合分数。考虑到分类和定位对特征需求存在一定的差异，在头部检测网络中引入特征对齐网络层，通过可形变卷积获得对齐后的特征，并与原始的回归特征组合成新的特

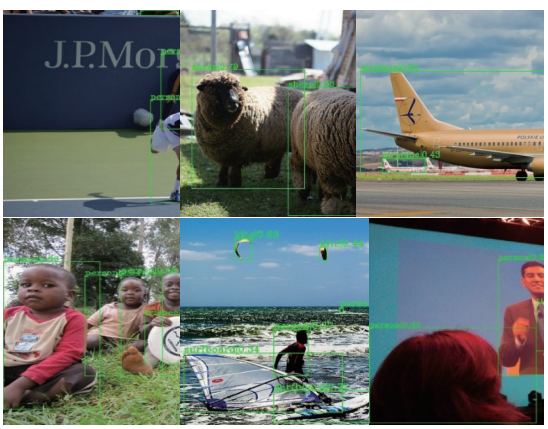

图4 端到端目标检测器检测结果可视化

征，用于回归边界框。基于特征对齐的头部检测网络缓解了分类和定位在特征层面的冲突，提升了端到端目标检测的性能。由于特征金字塔网络对目标特征提取存在一定的重复，网络输出仍有一定的冗余。如何进一步提高网络检测的稀疏性是我们下一步的工作重点。AP

作者简介：

王　宏（1972-），女，北京人，高级工程师，硕士，现就职于北京轩宇空间科技有限公司，研究方向为图像与视频处理、应用电子技术。

曾　峥（1984-），女，北京人，工程师，硕士，现就职于北京控制工程研究所，研究方向为图像处理、视频理解、射频通信系统。

徐奕男（1990-），男，北京人，工程师，硕士，现就职于北京轩宇空间科技有限公司，研究方向为图像信号处理、自动化控制。

谷晓琳（1992-），女，河北石家庄人，工程师，硕士，现就职于北京轩宇空间科技有限公司，研究方向为图像处理、计算机视觉与智能信息处理。

参考文献：

[1] 刘孙相与, 李贵涛, 詹亚锋, 等. 基于多阶运动参量的四旋翼无人机识别方法[J]. 自动化学报, 2022, 48 (6) : 1429 - 1447.
[2] 朱敏超, 冯涛, 张钰. 基于FD-SSD的遥感图像多目标检测方法[J]. 计算机应用与软件, 2019, 36 (1) : 232 - 238.
[3] Cai Z, Vasconcelos N. Cascade r-cnn: Delving into high quality object detection[C]. In: Proc. of the IEEE conference on Computer Vision And Pattern Recognition, 2018: 6154 – 6162.
[4] Girshick R. Fast r-cnn[C]. In: Proc. of the IEEE International Conference on Computer Vision, 2015: 1440 – 1448.
[5] He K, Gkioxari G, Doll´ar P, Girshick R. Mask r-cnn[C]. In: Proc. of the IEEE International Conference on Computer Vision, 2017: 2961 –2969.
[6] Ren S, He K, Girshick R, et.al. Faster R-CNN: Towards real-time object detection with region proposal networks[C]. In: Advances in Neural Information Processing Systems (NIPS), 2015.
[7] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]. In: European Conference on Computer Vision, 2016: 21 – 37.
[8] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]. Proc. of the IEEE International Conference on Computer Vision. 2017: 2980 - 2988.
[9] Tian Z, Shen C, Chen H, et al. Fcos: Fully convolutional one-stage object detection[C]. In: Proc. of the IEEE/CVF International Conference on Computer Vision, 2019: 9627 – 9636.
[10] Neubeck A, Van Gool L: Efficient non-maximum suppression[C]. In: 18th International Conference on Pattern Recognition (ICPR’06), 2006, 3: 850 – 855.
[11] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]. In: European Conference on Computer Vision, 2020: 213 – 229.
[12] Wang J, Song L, Li Z, et al. End-to-end object detection with fully convolutional network[C]. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2021.
[13] Sun P, Jiang Y, Xie E, et al. Onenet: Towards end-to-end one-stage object detection[C]. arXiv preprint arXiv: 2012, 057.
[14] Li X, Wang W, Wu L, et al. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection[C]. In: NeurIPS, 2020.
[15] Zhang H, Wang Y, Dayoub F, et al. Varifocalnet: An iou-aware dense object detector[C]. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2021.
[16] T. Lin, M. Maire, S. J. Belongie, L. D. Microsoft COCO: common objects in context[J]. CoRR, abs, 2014, 5.
[17] F. A. Group. Flir thermal dataset for algorithm training[DB/OL]. https://www.flir.in/oem/adas/adas-dataset-form/, 2018.
[18] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770 – 778.
[19] Chen K, Wang J, Pang J, et al. Mmdetection: Open mmlab detection toolbox and benchmark[J]. arXiv preprint arXiv, 2019.
[20] Deng J, Dong W, Socher R, et al. Imagenet: A largescale hierarchical image database[C]. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009: 248–255.
[21] Zhou X, Wang D, Krähenbühl P. Objects as points[J]. arXiv preprint arXiv, 2019.
[22] Krizhevsky A, Sutskever I, Hinton G.E. Imagenet classification with deep convolutional neural networks[J]. Advances in Neural Information Processing Systems, 2012: 1097–1105.