面向小目标和遮挡目标检测的脑启发 CIRA-DETR 全推理方法

宁欣[1),2),3)] 田伟娟[2)] 于丽娜[1)] 李卫军[1),2),3)*]

[1)]（中国科学院半导体研究所 高速电路与神经网络实验室，北京 100083）

[2)]（威富集团 形象认知计算联合实验室，北京 100083）

[3)]（中国科学院大学 集成电路学院，北京 100029）

摘 要 Facebook AI 研究者 2020 年提出的 Detection Transformer（DETR）目标检测方法采用简单的编码器-解码器结构，利用集合预测来解决物体检测问题，算法简单、通用、避免了很多手工设计和调参问题，吸引了学术界和产业界的广泛关注。然而，DETR 方法对于输入特征的分辨率大小有限制，同时在检测推理过程中缺失相对位置信息，从而导致对小目标和被遮挡目标的检测性能较差。为解决这一问题，受脑认知启发，本文提出基于胶囊推理和残差增强的全推理目标检测网络(Capsule-Inferenced and Residual-Augmented Detection Transformers, CIRA_DETR)。首先，建立层间残差信息增强模块，利用大小尺度的差异性对小尺度特征图进行信息增强，在小目标的检测效果上提升了 1.8%。接着，为了更贴近人脑的思维方式，更好的建模神经网络中内部知识表示的分层关系，在 Transformer 的结果进行推理的过程中，引入胶囊推理模块挖掘实体信息，并利用双向注意力路由进行前向信息传递和后向信息的反馈，以此预测图像中目标的类别和位置信息，有效降低了遮挡下的目标检测问题的难度。最后，在目标信息的映射处理中，引入非线性超香肠映射函数，实现了灵活的超曲面构建，有效表达特征和目标类别以及位置之间的映射关系。在 COCO 数据集上的测试结果验证了 CIRA_DETR 模型的有效性，其在小目标、中目标和大目标的检测上，平均预测准确率分别达到了 25.8%，48.7%和 62.7%。本文小目标的检测性能可以和 Faster-RCNN 相媲美，同时可视化的结果以及性能指标也反映了，相比传统的 DETR 模型，本文 CIRA_DETR 模型在被遮挡目标检测上的优势。

关键词 目标检测；DETR；Transformer；胶囊网络；脑神经科学；残差网络
中图法分类号 TP319

Brain-inspired CIRA-DETR full inference model for small and occluded object detection

NING Xin[1),2)] TIAN Weijuan[2)] YU Lina[1)] LI Weijun[1),2),3)]

[1)](Laboratory of Artificial Neural Networks and High-speed Circuits, Institute of Semiconductors, Chinese Academy of Sciences, Beijing 100083, China)

[2)](Image Cognitive Computing Joint Lab, Wave Group, Beijing 100083, China)

[3)](School of Integrated Circuits, University of Chinese Academy of Sciences, Beijing 100029, China)

Abstract In recent years, the deep learning has been applied in many image processing tasks. Deep learning shows an excellent performance and encourages many researchers to apply it to object recognition, including various popular directions: improve the direction accuracy by updating the network structure, design a simple network model based on the transformer, and obtain better detection results through the characteristic analysis of

本课题得到国家自然科学基金(No. 61901436)资助.

宁欣，博士，副研究员，硕士生导师， 计算机学会（CCF）高级会员（86216M）,主要研究领域为图像处理、模式识别. E-mail:ningxin@semi.ac.cn. **田伟娟**，硕士研究生，主要研究领域为神经元建模以及图像视频处理. E-mail: tianweijuan@wavewisdom-bj.com. **于丽娜**，博士，助理研究员，计算机学会（CCF）会员（J2616M）,主要研究领域为机器学习，深度建模及智能系统. E-mail: yulina@semi.ac.cn. **李卫军**（通信作者），博士，研究员，博士生导师，计算机学会（CCF）高级会员（95678S），主要研究领域为图像处理、模式识别. E-mail:wjli@semi.ac.cn

the gantry. The DETR object detection model proposed by Facebook AI researchers in 2020 utilize simple encoder-decoder structure, and views object detection as a direct set prediction problem. The DETR model is simple, general, and can avoid many manual designs and tuning problems, attracting widespread attention of the academia and industry. However, due to the limitation of Detr model on the size of input feature map, too small size will lead to insufficient object information. Although the performance of the model has been improved to a certain extent, its detection effect on small targets and occluded targets is not ideal. In the detection of small targets and occluded targets, the entity information corresponding to features and the relative position information between entities are very key to target reasoning. However, in Detr model, feedforward neural network FFN only realizes target information reasoning through weighted summation, and does not consider the interactive information between features, which has become the main factor affecting the detection effect. In contrast, humans can easily detect small targets and occluded targets. In order to solve these problems, we inspire by brain cognition, propose a novel full-inference model, called Capsule-inference and residual-augmented DETR (CIRA_DETR) inspired by the brain cognition. Firstly, CIRA_DETR establishes an inter layer residual information enhancement module to enhance the target related information in the small-scale map by calculating the differences between the large and small-scale feature maps, which can improve the convergence speed and detection performance of Detr model without increasing the complexity of the algorithm. Thus, the performance for small object detection improves by 1.8%. Then, in order to more closely match the way of human brain thinks and to better model the hierarchical relationships of intra-knowledge representation in the neural networks, during the inference process, capsule-inference module is constructed to explore the object entity information, and utilize the bi-directional attention routing for forward information delivery and backward information feedback, and then the object class and location information can be inferenced. With the capsule-inference module, the problems for occlude object detection has been greatly alleviated. Finally, neuroscience suggests that anything human sees converges to a continuous attractor in different ways, which may be a curve, a surface, or a hypersurface. Inspired by this brain science hypothesis of continuous attractor, during the mapping process for object information, a hypersausage measurement model with stronger nonlinear ability is introduced to form a more flexible hypersurface, so as to describe the mapping relationship between features and labels and improve the expression ability of the model for the target. In the experiment of MS COCO dataset, the object detection precision respectively achieves 25.8%, 48.7% and 62.7% for small, middle and large objects. Extensive experiments conducted on the representative MS COCO dataset show the effectiveness of the proposed CIRA DETR model, especially for object occlusion and small object detection.

Key words Object detection; DETR; Transformer; CapsNet; Neuroscience; ResNet

1 引言

目标检测任务旨在为每个预先确定的物体预测一组边界框和类别标签[1]。两阶段目标检测算法[2, 3, 4, 5, 6, 7]主要通过生成大量的区域建议、预测每个候选框、并利用非极大值抑制策略去消除高度重叠的候选框来实现目标检测，这类方法由于复杂的计算流程使得其很难部署和投入使用。Carion 等[8, 9, 10]提出了一种 Detection Transformer（DETR）方法，它基于编码器-解码器框架并结合集合预测的直观方式来解决目标检测任务。DETR 使用一个深度残差网络（ResNet）[11]主干来提取特征，结合位置编码，将其传递到编码器中；然后，解码器将学到的少量固定数量的位置嵌入作为对象查询向量，并将其作为额外的注意力添加到编码器的最后一个迭代层；最后，采用匈牙利匹配损失将解码器的输出归一化之后回传到前馈网络（Feed Forward Network，FFN），FFN 负责检测目标（类别和边界框）或 “无物体”类。DETR 通过利用匈牙利匹配算法进行预测目标框和标注框的匹配进行模型训练。

借助 Transformer[12]的学习能力，DETR 可以进行合集预测，而无需锚点设计和区域建议等人工先

验，因而被认为是一种更简单的目标检测框架。为获得良好的性能，DETR 需要输入高分辨率图像，使得编码器和解码器提取所有位置交互信息时，计算量翻倍，存在编码器计算复杂度高和收敛速度慢的问题。为此，很多研究工作致力于改进 DETR 方法[13-22]以降低计算复杂度、提高收敛速度。如：Zheng M.等[13]分析了编码层中输入特征的相似性并将其可视化，发现相似的点在注意力图中也是相似的，基于此，他们提出使用局部敏感哈希（Locality Sensitive Hashing，LSH）来进行查询特征的自适应聚类方法（Adaptive Clustering Transformer，ACT），替代了 DETR 中的自注意模块，在不影响预训练性能的情况下，将自注意力的时间复杂度 O(N2)降低到 O(NK)，在准确性和计算成本之间实现了良好的平衡；此外，Zhu X.等[14]提出基于可变形注意力模块的 DETR 框架，仅需关注参考点周围少量的关键采样点，使得训练 epoch 数目减少了 10 倍，收敛速度更快；进一步，Sun Z.等[15]指出 DETR 收敛慢的主要因素在于二分匹配部分的不稳定性和 cross attention 稀疏度问题，进而提出了以 TSP-FCOS 和 TSP-RCNN 模型作为解码器结合 DETR 编码器的目标检测方法，有效提高了收敛速度。然而，由于 DETR 模型对输入特征图的大小有限制，太小的尺寸将导致对象信息不足，虽然模型性能得到了一定程度的提升，但其对小目标和遮挡目标的检测效果却并不理想。

在小目标和遮挡目标的检测任务中，与特征相对应的实体信息，以及实体之间的相对位置信息对于目标推理是非常关键的。然而，在 DETR 模型中，前馈神经网络 FFN 仅通过加权求和实现目标信息推理，并未考虑到特征间的交互信息，成为影响检测效果的主要因素。反观人类，可以很容易的检测到小目标和遮挡目标。脑神经科学认为，人类看到的任何东西都会以不同的方式汇聚到一个连续吸引子，该吸引子可能是一条曲线、一个曲面或者超曲面[16]。基于此假设，可以在在高维空间中通过连续的拓扑几何体对样本进行覆盖学习[17]。此外，Hinton 等[18]以人脑的思维方式出发，构建胶囊实现实体及其属性的描述，并利用动态路由信息传递方式，更好的建模网络中内部知识表示的分层关系和传递方式。对于机器目标检测任务，这种超曲面的设计将对检测效果产生重要影响。在传统模型中，为了刻画特征与目标类别、位置信息之间的关系，会使用一个带有核函数的非线性超曲面进行映射，如 M-P[19]神经元模型、径向基函数（Radial Basis Function，RBF）[20]、生物拟真神经元模型 Flexible Transmitter（FT）[21]等，这些模型虽在一定程度实现了超曲面的构建，但却无法兼顾超曲面所在空间中的方向、位置和区域大小三个维度，非线性刻画能力较弱。

基于上述分析，本文受人类脑神经科学启发，将图片的目标检测任务转换成：“图片中有几个目标，每个目标的所属类别是什么，每个目标在图片中的位置是哪里”的推理任务，提出基于胶囊推理和残差增强的全推理目标检测网络(Capsule-Inferenced and Residual-Augmented Detection Transformers, CIRA_DETR)，以一维单纯形作为基本单元，构建具备灵活形状的超香肠超曲面来增强模型的表达能力，并基于胶囊网络构建思想，采用动态路由信息传递方式，利用特征之间的相关性实现目标表示和标签预测，建立了全推理式的目标检测方法，可以有效完成小目标和遮挡目标检测任务。具体创新点如下：

1）建立层间残差信息增强模块，通过计算大小尺度特征图之间的差异性，增强小尺度图中目标相关信息，可实现在不增加算法复杂度的情况下，从而提高 DETR 模型的收敛速度和检测性能。

2）为了贴近人脑的思维方式，模拟网络中内部知识表示的分层和传递方式，构建胶囊推理模块，结合胶囊构建和信息路由方式，进行目标类别和位置的推理，该胶囊推理模块能够在推理的过程中，引入目标的相对位置信息和类别关联，能够捕捉全局信息，增强目标的检测性能。

3）基于事物最终会汇聚到一个连续吸引子的脑科学启发，在模型的推理过程中，引入一种非线性能力更强的超香肠度量模型，用于形成更灵活多变的超曲面，以此刻画特征和标签之间的映射关系，提高模型对于目标的表达能力。

4）在 COCO 数据集上的实验证明了所提出的几个模块和 CIRA_DETR 在实现高性能方面的有效性，特别是在小物体和遮挡物体检测方面。

2 相关工作

2.1 小目标检测

尽管在许多应用中已经实现了对图像中大中型物体的精确检测，但对小物体的精确检测，仍然具有挑战性。由于特征难以区分、分辨率低、背景

复杂、上下文信息有限等原因，小目标难以检测。近年来，针对小目标的检测有很多研究。例如，Kisantal M.等[22]提出过采样和复制粘贴小对象来解决这一问题。感知 GAN[23]通过生成超分辨特征，并将其堆叠到小物体的特征映射中以增强表示。DetNet[24]保持了空间分辨率，并具有较大的接收域，以提高小目标检测。SNIP[25]将图像调整到不同的分辨率，只训练接近地面真相的样本。YOLO-CAN[26]通过在特征提取网络的每个残差块的通道和空间维度上引入注意机制，对小目标进行聚焦。MTGAN[27]是一种端到端多任务生成对抗网络，生成器是一个超分辨率网络，它可以将模糊的小图像上采样成小尺度的图像，并恢复出细节信息，从而实现更精确的检测。MSFYOLO[28]则通过实现具体特征与抽象特征的融合来实现对小目标的检测。与这些方法不同的是，我们通过建立层间残差信息增强模块，利用大小尺度的差异性对小尺度特征图进行信息增强，不仅有效地提高了小目标的检测性能，而且保证了其他目标的检测性能。

2.2 遮挡目标检测

遮挡问题是目标检测任务中另一个挑战性问题，主要包括以下两类，一方面是检测目标间存在的相互遮挡，另一方面是待检测目标被干扰物体遮挡。例如，在行人检测应用中，在拥挤的街道、火车站和工厂中普遍存在遮挡现象，遮挡下的行人图像呈现出各种形状和形式。在处理变形和遮挡问题时，行人检测算法的精度会降低[29]。从部件优化的角度，Tian Y.等[30]提出了一个由 45 个不同组件组成的组件池，将每个分量的卷积神经网络训练为子检测网络，最后通过对所有子检测器综合计算得分来表征最终的检测结果。Ouyang W.等[31]提出了一种 DP-CNN 方法，使用模式挖掘算法将目标的各种局部信息提取出来，然后利用这些特征训练局部特征检测器。最后，将局部特征检测器嵌入到卷积神经网络中，提高目标检测算法面向遮挡问题的处理能力。从损失函数改进角度，Wang X.等[32]提出了一种基于吸引、排斥项的损失函数。。该损失函数通过使预测帧接近匹配的实帧而远离不匹配的实帧来控制重叠部分的面积。YOLO-CAN[26]将高斯模型用于非极大抑制损失函数，以增强对遮挡目标的检测能力。IA[33]利用内部干扰训练的目标检测网络将注意力分散到整个目标上。从而有效地提高了训练中特征的多样性，使网络对图像缺陷具有鲁棒性。内部和内部的共现信息都有助于改善特征表示，以处理不同水平的遮挡。近年来虽然遮挡问题的研究取得了一定进展，但存在时间复杂度过高、针对遮挡问题的优化效果不足等相关的问题。

2.3 胶囊网络

作为“仿生派”的代表人物 Hinton,他提出的“胶囊网络（CapsNet）[34]是专门为基于 CNN 的特征提取而设计的，已经得到了人工智能研究人员的极大关注。胶囊网络的传输和运算逻辑更符合人脑神经元的工作方式，不同的胶囊可以携带不同属性，就像人脑的不同区域负责不同的工作。随着神经科学对人脑认识的不断深入和持续积累，脑科学和人工智能的研究开始融合。一个胶囊是一组神经元，其活动向量代表一个具有特定类型的实体的实例化参数。与传统的 CNN 中特征由特征图中标量值表示的方法不同，CapsNet 中的特征是由胶囊表示的。胶囊的方向反映了特征的特性，胶囊的长度反映了不同特征存在的可能性。层与层之间的信息传递遵循动态路由机制，将低层模块的全部信息路由到最接近的高层模块，这使得高层特征的生成成为可能。此外，CapsNet 使用多个嵌入进行图的建模，每个嵌入揭示了图的不同属性，相比其他基于标量仅使用一个嵌入的方法，CapsNet 具有更好的图表示能力。

CapsNet 模型利用胶囊输出向量的长度来表示该胶囊所代表的实体在当前输入中存在的概率，并采用非线性“squash”函数来确保短向量被缩减到几乎为零的长度，而长向量则被缩减到略低于 1 的长度，从而实现了判别性的学习。

$$\mathbf{v}_j = \frac{\|\mathbf{s}_j\|^2}{1+\|\mathbf{s}_j\|^2}\frac{\mathbf{s}_j}{\|\mathbf{s}_j\|} \qquad (1)$$

其中，$\mathbf{v}_j$ 是胶囊 j 的矢量输出，$\mathbf{s}_j$ 是输入。$\mathbf{v}_j$ 的方向反映了胶囊的特性，而 $\mathbf{v}_j$ 的长度反映了胶囊存在的可能性。CapsNet 的提出受到了人工智能研究者的极大关注。

针对 CapsNet 模型基于迭代算法的动态路由[18]所存在的一些问题，Hinton G.E.等[35]利用 EM 路由创建矩阵胶囊，将实体表现为一个姿势矩阵。进一步，Wang D.等[36]将动态路由作为一个优化问题，引入耦合分布之间的 Kullback-Leibler（KL）散度进行求解。此外，Jaiswal A.等[37]将胶囊网络成功用到 GAN 网络中，将其作为判别器建立 Capsulegan 方法，获得了比传统 GAN 更好的视觉性能。LaLonde R.等[38]提出 SegCaps 方法，成功将胶囊网

络应用到图像分割中，并在 LUNA16 数据集中取得了很好的效果。

上述分析可知，胶囊网络对于实体信息的挖掘、实体之间的交互以及对于目标的推理方面，已经取得了较好的性能。在目标检测任务中，如何在有限的条件域中检测隐式定义的实体，并实时刻画实体的特征信息（如：实体的位置、类别和姿态信息等）非常关键。因此，胶囊网络预期可在目标检测任务中发挥重要作用。

3 方法

CIRA_DETR 方法的总体框架如图 1 所示。首先，输入图片经过 ResNet 主干网络提取特征，将所获取的不同尺度的特征图经过本文提出的层间残差信息增强方法，用于实现小尺度信息的增强，增强后的特征图如图 1 中的红框标注所示，对应于原图中的两个高尔夫球小目标；接着将增强后的特征图输入到 Transformer 的编码器和解码器中，用于获取各个目标的特征表示；最后将各个目标对应的特征图输入到本文新提出的基于超香肠度量的胶囊推理模块，用于构建胶囊实体，并进行信息路由向上层进行传递，实现目标类别和位置信息的挖掘。此外，为了刻画胶囊实体和标签之间的映射关系，引入一种非线性超香肠度量模型，用于映射得到目标的位置和类别结果。CIRA_DETR 模型综合考虑了特征层间的交互信息和实体间的相对位置信息，能够有效提高目标推理的性能。下面我们分 3 个小节详细描述本文所提出的残差信息增强模块，胶囊推理模块以及香肠度量模块。

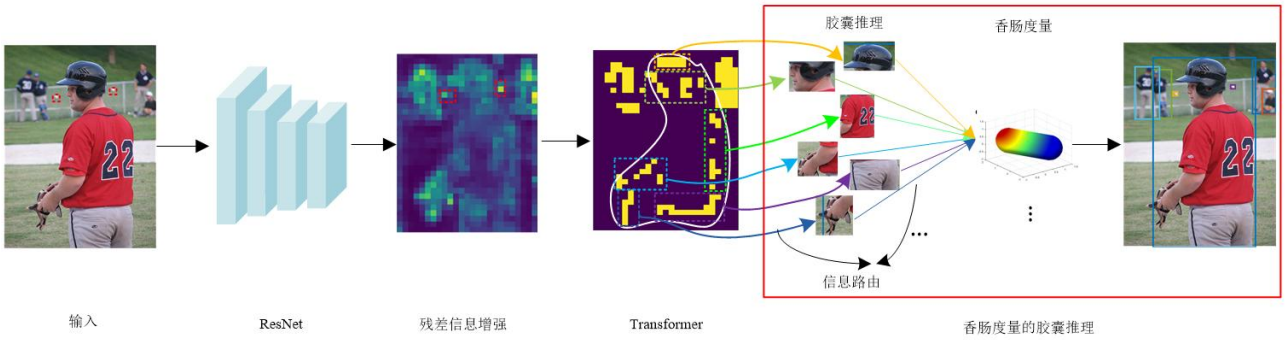

图 1 基于胶囊推理和残差增强的全推理目标检测网络(CIRA_DETR)结构

3.1 残差信息增强

DETR 中，输入特征图太大可能会增加模型训练的难度，太小可能会导致特征图包含的目标信息缺失。为了充分利用大尺度所包含的特征图信息的同时，避免增加模型训练的负担，本文提出一种层间残差信息增强方法，将目标特征信息尽可能保留下来，以达到提高目标检测性能的目的。该模块主要通过挖掘相邻大尺度特征图和小尺度特征图之间的差异，并利用该差异性实现小尺度特征图信息的增强，能够检测出 DETR 所遗漏的小尺度目标，由于本文利用的大特征图的尺度是有限的，而小目标检测性能的提升是依赖于大尺度特征图的尺寸的，因此，小目标的尺寸越大，小目标检测性能越明显。具体如下：

1）将输入图像经过骨干(Backbone)网络，得到最后两层的特征图，分别记为 F_1 和 F_2，其中倒数第二层为 F_1，倒数第一层为 F_2。F_1 相对于 F_2 来说，包含较多的目标信息，而 F_2 则包含关于目标更多的语义信息，但也损失了部分信息，尤其是小目标的信息。

2）我们对 F_2 进行上采样和卷积处理之后，相当于利用 F_2 去重构 F_1 特征图，最后计算重构之后的特征图和 F_1 的差值结果，那么得到的便可合理的认为是丢失的目标信息，计算如式（2）

$$mask = Sigmoid(F_1 - Conv_{1\times1}(UpSampling(F_2))) \quad (2)$$

3）利用通用的处理方式将丢失信息的值，即将 $mask$ 附加到 F_1 上，从而实现了 F_1 层小目标信息的增强，再通过基于注意力的池化方法 PMA[39]算法，得到增强之后的结果 RA，计算如式（3），

$$RA(F_1, F_2) = PMA(F_1 * mask) \quad (3)$$

其中，$*$ 代表矩阵的点乘操作，经过残差信息增强之后的结果 RA，其对应的特征图的大小是等于 F_2 特征图的大小的，也就是说我们在未增加特征图尺寸的情况下，增强了小目标在特征图中的信息，将增强后的目标特征图用于后面的 Transformer 结构中，能够有助于获取较好的小目标检测效果。

3.2 胶囊推理

本文在胶囊网络的启发下，构建胶囊推理模

块，并成功应用于目标检测中，替换了 DETR 模型中用于推理的 FFN 模块。其实现方式主要是通过将目标的类别和坐标转化为胶囊中实体的属性表示，实体即为 Transformer 得到的目标表示。图 2 给出了胶囊推理模块用于实现目标检测的原理图。对于 Transformer 的输出，构建基础胶囊，其中包含了胶囊及对应的特征维度信息，接着对所有的基础胶囊，引入注意力信息传递的方式[40]，进行目标类别和位置相关信息的预测，得到推理胶囊。具体实现过程如下：

1）将图像输入到 ResNet 网络中，并经过残差信息增强模块，得到增强后的特征图 RA，再输入到 Transformer 的编码器和解码器中，为每一张图像得到关于 100 个目标的特征表示；

2）将上述得到的特征表示进行胶囊转换，转化之后胶囊大小变为 $[bs, 100, head, dim]$，其中 $head$ 表示胶囊的个数，而 dim 表示每个胶囊的向量表示，用于表示胶囊的姿态、纹理、方向等等，至此，得到基础胶囊表示 $PCaps$；

3）以 $PCaps$ 特征图作为下层，通过注意力路由传递方式，去推断出图片当中所包含的目标信息，包括：图片当中每个目标的所属类别是什么，以及每个目标在图片中的位置是什么，大小为 $[bs, 100, classes+coord]$，其中 $classes$ 代表目标类别个数，值为 92，每个值代表在每个类别下的置信度。$coord$ 代表目标框的坐标数目，值为 4，表示为目标在图像中的位置信息。至此，得到数字胶囊表示 $DCaps$。

4）根据所得到的 $DCaps$，选择各个目标置信度对应的得分最高值且值大于 0.9 的所对应的类别即为该目标所属类别，对应的坐标即为其在图像中的位置信息。至此，得到当前图像的目标检测结果。

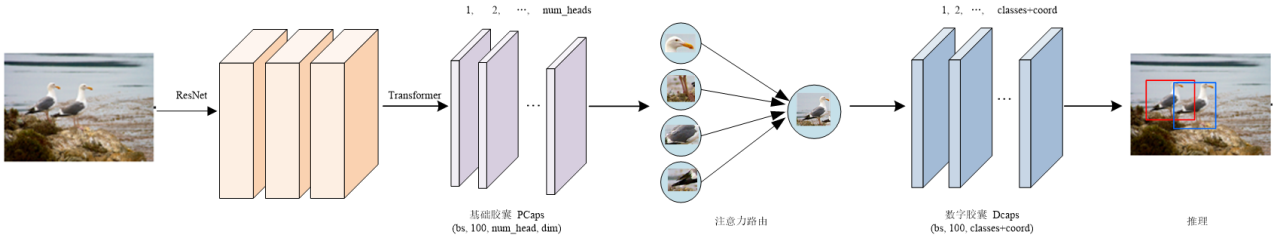

图 2 CIRA_DETR 模型中基于胶囊推理模块的目标检测实现原理

3.3 基于脑启发的香肠度量

神经科学与认知科学往往对于神经网络模型设计具有重要启发作用。一方面，在功能机制方面，神经科学研究表明，如果用于记忆的离散吸引子能被替换成了连续吸引子，视觉信息在人脑中不是一个点，而是一条曲线、一个曲面甚至超曲面[16]。根据连续吸引子的假设，可以合理地假设，如果有两个“同源”而相似的但不完全相同的事物，那么在这两个相似的事物之间至少有一个渐进的过程，并且在这个过程中的所有事物都属于同一类，称为同源连续性原理[41-42]。基于同源连续性原理，对一种事物的“认识”，实质是对这种事物在特征空间中的全体形成的无限点的集合的“形状”的分析和“认识”[43]。另一方面，在结构层面之上，大脑不同部位的神经细胞以不同的方式处理信息，具有不同的功能，不同类型的脑神经元在整个视觉皮层形成一个功能结构，用于合成和提取信息[44]。受上述脑认知的启发，本文旨在构建一个形状丰富、能够描绘目标不同角度和尺寸的双曲线神经元，基于此，考虑到胶囊输出结果和目标标签之间较为复杂的映射关系，我们进一步扩展了神经元的构造方法和学习范式，并引入了一种非线性能力较强的超香肠判别函数，用于完成特征和标签之间的映射关系，该函数来源于对 RBF[20]函数的扩展，将 RBF 的中心扩展成一条线段，并沿着固定半径去旋转，得到类似超香肠的几何体，相对于 RBF 来说，其具备更高的自由度，可提高模型的判别能力。

推理胶囊的输出向量中，我们以每个目标作为胶囊，那么目标的类别和坐标即对应目标胶囊的输出维度值，传统的胶囊模型采用 squansh 激活函数，计算每个目标胶囊各个属性在目标中出现的概率值作为判断目标类别的依据。图 3 给出了超香肠神经元模型在 DETR 中的应用流程，主要是将数字胶囊的输出输入到超香肠度量模型中，去计算对应到每个类别的概率值以及对应类别下的位置坐标，然后根据选择概率最大的值为作为每个目标的类别，获取对应的位置坐标作为检测结果。

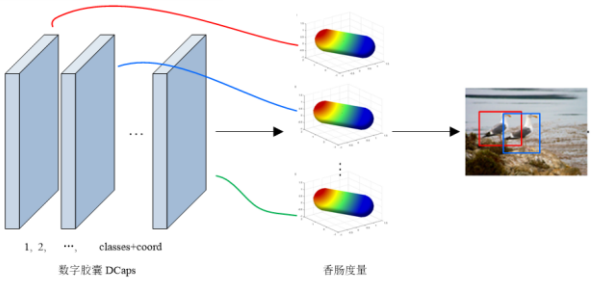

图 3 CIRA_DETR 模型中基于香肠度量的目标检测实现原理

超香肠度量模型的构建方式主要是将单个向量扩张到半径为 r 的一个香肠区域中，通过计算样本点距离香肠模型的距离，得到样本点和该香肠模型之间的关系，定义如下：

$$y=\Phi\left(\frac{\|x-[\lambda\mathbf{q}_1+(1-\lambda)\mathbf{q}_2]\|^2}{n\cdot r^2}\right) \qquad (4)$$

其中，$y\in[0,1]$ 是预测的第 i 个神经元输出；x 是输入向量；r 是超香肠度量的半径；$\mathbf{q}_1$ 和 $\mathbf{q}_2$ 是端点，记为香肠的核。n 是为了平衡高维的稀疏性而增加的特征维度；Φ 表示激活函数，这里采用高斯核激活。λ 表示 $\mathbf{q}_1x$ 在 $\mathbf{q}_1\mathbf{q}_2$ 上的投影长度，表示为：

$$\lambda=\begin{cases}1, & \mathbf{k}(\mathbf{q}_2-\mathbf{q}_1)<0\\ \|\mathbf{k}\|, & \dfrac{\|\mathbf{k}\|}{\|\mathbf{q}_2\mathbf{q}_1\|}<1,\ \mathbf{k}(\mathbf{q}_2-\mathbf{q}_1)\geq 0\\ 0, & \dfrac{\|\mathbf{k}\|}{\|\mathbf{q}_2\mathbf{q}_1\|}\geq 1,\quad \mathbf{k}(\mathbf{q}_2-\mathbf{q}_1)>0\end{cases} \qquad (5)$$

其中 $\mathbf{k}$ 是 $\mathbf{q}_1x$ 在 $\mathbf{q}_1\mathbf{q}_2$ 的投影向量，计算如下：

$$\mathbf{k}=\frac{(x-\mathbf{q}_1)(\mathbf{q}_2-\mathbf{q}_1)}{\|\mathbf{q}_2-\mathbf{q}_1\|} \qquad (6)$$

$d=\|x-(\mathbf{q}_1+(1-\lambda)\mathbf{q}_2)\|$ 是输入 x 到矢量 $\mathbf{q}_1\mathbf{q}_2$ 的距离。如果 $d/r>1$，相应的样本就在超香肠几何体之外，被判断为负样本，否则，相应的样本就在超香肠几何体之内，是正样本。分布在不同位置的特征点的具体距离计算如图 4 所示，其中 d 代表样本点与矢量的距离，形状相同的特征点符合相同的距离计算方法。

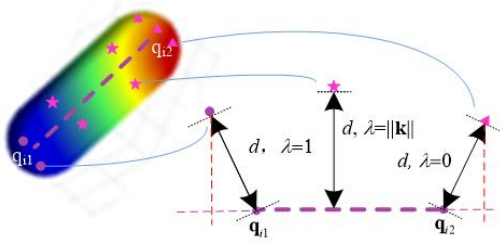

图 4 特征点与香肠区域之间的距离示意图。圆圈、三角形和星星分别代表分布在香肠区域不同位置的样本

4 实验

4.1 数据集

本文在 COCO 2017 检测数据集[45]上进行了实验，该数据集包含 118k 训练图像和 5k 验证图像。数据集中的每张图像包含多达 63 个不同大小的实例。我们将 AP 报告为 bbox AP，即多个阈值的积分指标。我们还报告了 COCO 2017 验证集中前 100 张图像的平均 FLOPs。只有卷积层、全连接层、矩阵运算的 FLOPs 在注意力上会被考虑。

4.2 实验设置

考虑到 DETR[46]模型的学习对于显卡的要求，网络从头训练是非常吃力的，因此，本文所有的实验都是采用 DETR 预训练的模型进行训练的。在本文 CIRA_DETR 模型的训练过程中，我们固定 backbone 中 layer2 之前的所有权重，针对后面的 backbone 层，增大学习率至 1e-4，后面编码器、解码器，设置学习率为 1e-5 进行微调，推理层设置学习率为 1e-4。在实现 CI_DETR 模型的消融实验学习中，固定 backbone、编码器和解码器模块的权重参数，只学习推理层的参数。在实现 RA_DETR 模型的消融实验中，学习率设置和 CIRA_DETR 的学习保持一致。DETR 包含了 6 个编码层和 6 个解码层，每层注意力有 8 个 head。Primary 胶囊层的胶囊数目设置为 8，Inference 胶囊层的胶囊数目设置为 100。考虑到 DETR 模型对 GPU 显存的需求，从头开始训练网络是非常困难的，因此，本文的所有实验都是使用预训练的 DETR 模型进行模型微调，并将学习率降低 1/10，采用 AdamW[47, 48]，优化器进行优化，三张 Quadro RTX 5000 GPU 显卡被用于训练和测试。

在本文 CIRA_DETR 模型的训练过程中，我们固定了 backbone 倒数第 3 层之前的所有权重，将 backbone 层的学习率提高到 1e-4，将后面的编码器和解码器的学习率设置为 1e-5，将推理层的学习率设置为 1e-4。在实施胶囊推理模型学习的消融实验中，backbone 层、编码器和解码器模块的权重参数是固定的，只学习推理层的参数。而在消融实验中，为了实现残差增强模型的学习，学习率设置与 CIRA_DETR 学习一致。

4.3 实验结果与分析

为了验证所提出的 CIRA_DETR 模型相对于传统的基于 anchors 和 anchor-free 的目标检测方法的有效性和优势，表 1 展示了我们在 COCO2017 验证集上的主要结果，包括 AP（Average precision）[47]值以及 FPS 值。我们将 CIRA_DETR 与经典网络，如 FCOS[49], Faster RCNN[6],YOLOv3[50]，改进的 DETR[13,14]，以及损失改进模型[46,47]进行比较。

从表 1 中，我们可以看到，本文提出来的 CIRA DETR 的表现明显优于对比模型。为了与最先进的 DETR 模型进行比较，我们使用了在 DETR 中类似的训练策略[1]。其中采用了 96 个周期（8*）的训练计划和随机裁剪数据增强策略。对比这些模型，我们可以发现本文 CIRA_DETR 与经典的 FCOS[22]、Faster RCNN[6]和 YOLOv3[50]模型相比，取得了较大的提升幅度。同时，与解决小目标与遮挡问题的 MFSYOLO[28]、YOLO-ACN[26]和 IA[33]方法相比也取得了更好的结果。此外，和 DETR 以及改进后的模型相比，也取得了一定的性能提升。此外，相对于损失函数改进模型，本文中的 CIRA DETR 也显示出巨大的优势。

表 1 CIRA_DETR 模型在 COCO 2017 验证集的评估结果

Model	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	FLOPs	FPS
FCOS†[49]	ResNet-50	39.9	59.1	43.4	23.2	42.6	50.9	177G	17
Faster RCNN[6]	ResNet-50	40.2	61.0	43.8	24.2	43.5	52.0	180G	19
YOLOv3[50]	DarkNet-53	33.0	57.9	34.4	18.3	25.4	41.9	30.2G	20
MFSYOLO[28]	ResNet-101	33.5	52.9	34.9	23.6	44.1	48.5	27.7G	41.1
DETR+[46]	ResNet-50	41.9	62.3	44.2	20.3	45.8	61.0	86G	21
DeformCaps[14]	DLA-34	40.6	58.6	43.9	23.0	42.9	56.4	---	15
ACT[13]	ResNet-50	42.6	61.2	44.1	21.4	46.8	61.1	58G	15
YOLO-ACN [26]	DarkNet-53	34.1	58.9	34.6	20.5	28.5	45.7	47G	22
CornerNet[51]	Hourglass-104	40.6	56.4	43.2	19.1	42.8	54.3	---	---
RetinaNet[52]	ResNet-50	38.7	58.0	41.5	21.0	42.3	50.0	---	24
IA[33]	ResNet-101	41.0	62.0	45.1	23.6	45.8	52.0	104G	25
CIRA_DETR	ResNet-50	**43.0**	**63.3**	**45.5**	**24.6**	**46.8**	**62.1**	90G	19

†代表复现结果；+代表模型是用随机此裁剪策略和较长训练策略的结果。

4.4 消融实验

本文层间残差信息增强、胶囊推理和香肠判别模块分别对所提出的 CIRA_DETR 模型的性能做出了特有的贡献。在消融分析中，我们进行了三个消融实验的研究，以探索三个基本模块如何影响我们的 CIRA_DETR 模型的结果。

4.4.1 残差信息增强

本文提出的残差信息增强方法，通过挖掘大尺度特征图和小尺度特征图之间的差异，并利用该差异性增强针对目标的小尺度特征图的表述能力，进而提升目标检测的性能。为了验证本文残差信息增强模块的有效性，表 2 给出了残差信息增强应用在 DETR 上的消融实验结果，记为 RA_DETR。由表 2 可以看出，在 DETR 模型中加入残差信息增强模块后，在模型复杂度并没有增加的基础上，性能有了一定幅度的提升，其在小目标上的检测性能可与 Faster RCNN 相媲美，也验证了残差信息增强对于所提出的 CIRA_DETR 模型的重要性。

表 2. 用于残余信息增强模块的消融研究在 COCO2017 验证集上的评价结果(%)

Model	Backbone	AP	AP_S	FPS
DETR+[46]	ResNet-50	42.0	20.3	**21**
ACT[13]	ResNet-50	42.6	21.4	15
Faster RCNN[6]	ResNet-50	40.2	24.2	19
RA_DETR	ResNet-50	**42.2**	**24.3**	20

为了进一步获得公平的实验结果，我们观察特征图中存在的目标信息，特别是包括小目标。具体来说，我们以小目标为例，观察特征图的变化，以及加入残余信息增强模块后的检测结果，结果见图 5。其中红色框标注的为小目标所在的区域及其对应的特征图结果。为清晰的查看小目

标的检测效果，小目标所在区域被放大，图像中其他的目标所在区域以及检测结果被忽略。从图 5 可以很明显的看出，使用本文提出来的 RA_DETR 模型，图 5(上)中 4 个小目标检测出来了 2 个，图 5（下）中的 2 个小目标均可正常检测，而 DETR 模型对于以上小目标显得无能为力。同时，增强之后的特征图在对应的小目标所在位置，其特征信息更加明显，即针对小目标的描述更加明确，有助于提升后期小目标的推理和检测性能。上述结果表明了，本文残差信息增强模块的引入，可以检测出普通的 DETR 模型遗漏的小目标，增强了小目标的特征信息，减轻了 DETR 小目标检测能力较差的问题。结果更充分的展示出本文提出来的 CIRA_DETR 模型的有效性。

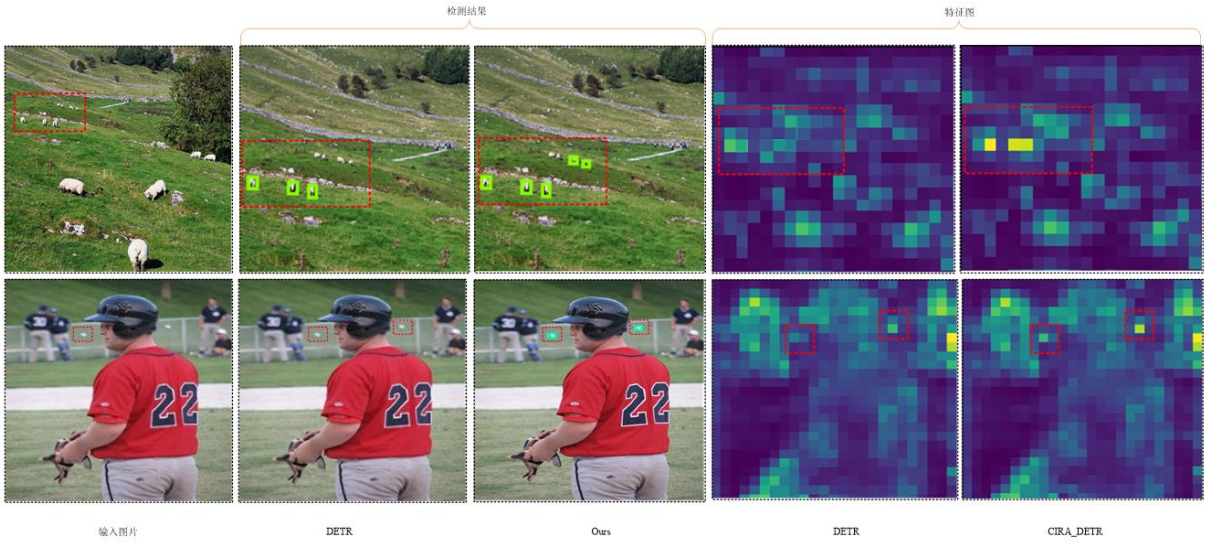

图 5 使用 DETR 和 CIRA_DETR 模型对图像中的小目标特征图进行可视化的示例

4.4.2 胶囊推理模型

本文构建的推理胶囊模块，利用胶囊构建和注意力路由传递的方式实现目标所属类别和位置信息的推理，避免了直接采用 FFN 模型时对于描述目标相对位置时的缺陷，增加了胶囊实体以及路由信息，用于实现模型的自主推理能力，进而提高目标的检测性能。表 3 给出了胶囊推理模块用于 DETR 的消融实验结果，记为 CI_DETR。从表 3 可以看出，胶囊推理模块用于 DETR 的准确率达到了 42.4%，明显高于其他 DETR 及其改进模型，表明了本文的胶囊推理模块对于 CIRA_DETR 的重要性，其原因在于胶囊推理模型通过构造胶囊实体，可以实现对于目标更全面的内部表示，并根据胶囊实体之间的路由信息传递方式，可以获取胶囊实体对于目标更有效的推理方式。

表 3. 胶囊推理模块在 COCO2017 验证集的评价结果

Model	Backbone	AP	AP_{50}	AP_{75}	FPS
DETR+[46]	ResNet50	42.0	62.4	44.2	21
DeformCaps[14]	DLA-34	40.6	58.6	43.9	15
RetinaNet[52]	ResNet50	38.7	58.0	41.5	24
CI_DETR	ResNet50	**42.4**	**62.8**	**44.9**	19

为了进一步验证胶囊推理模块面向目标检测任务的有效性，我们筛选有遮挡的图像，进行目标检测，图 6 可视化了有遮挡的图像经过本文 CI_DETR 和普通 DETR 模型之后的目标检测结果。第二列和第三列分别代表采用 DETR 和本文 CI_DETR 模型的检测结果，虚线红框标记的目标代表的是未被检测到的被遮挡的目标。由图 6 可以看出，第二列采用 DETR 算法只能检测出遮挡的目标，而本文提出来的胶囊模块能够检测出被遮挡的目标，如图 6 中被遮挡的鸽子、汽车、映在汽车表面的行人（下），但由于遮挡程度和像素过于模糊的问题，检测的目标个数和定位的准确度出现了一定的偏差。上述表明了胶囊推理模型在遮挡目标检测中的巨大优势。其原因在于，部分重叠的目标其特征容易混淆，导致目标类别和位置信息的预测发生偏差。若采用胶囊进行目标信息推理，由于每个胶囊实体代表目标的不同部件，能够提取并利用有关对象相对位置的信息内容，该信息内容可以解析重叠对象。同时，胶囊之间自下而上的信息传递可以充分利用胶囊实体独立的信息推理能力，挖掘每个胶囊实体对于整个目标检测的贡献，并最终实现检测性能的提升。

图 6 使用 DETR 和 CIRA_DETR 模型对图像中被遮挡物体的检测效果示例

4.4.3 香肠度量模型

本文在胶囊推理模块中，引入超香肠度量模型，用于获取每个目标胶囊各个属性在目标中出现的概率，该度量函数具备较强的非线性能力，能够很好的刻画底层胶囊（特征）和高层胶囊（类别和位置）之间的映射关系，进而提高特征的表达能力。表 4 给出了香肠度量模块用于 CIRA_DETR 模型中的消融实验对比结果。从表 4 我们可以看出，当替换胶囊中的 squash 激活函数为本文的超香肠度量模型时，模型的检测准确率有了较大幅度的提高。其原因在于，传统的胶囊构建函数中，采用 squansh 激活函数获取属性在目标中出现的概率，但其并未考虑到特征和类别之间复杂的映射关系，尤其在检测、分割等较为复杂的任务中。而本文所引入的超香肠度量模型即能满足属性和类别之间的映射关系，同时其灵活的非线性映射能力也有助于减轻复杂任务的识别难度。检测结果越好，也就证明了香肠度量模块对于检测性能的重要性和有效性。

表 4. 香肠测量模块的消融实验研究在 COCO2017 验证集上的评价结果

Backbone	香肠度量	AP	AP_S	AP_M	AP_L
ResNet50	是	**43.0**	**24.6**	**46.8**	**62.1**
	否	42.7	24.4	46.4	61.8
ResNet101	是	**44.5**	**25.8**	**48.7**	**62.7**
	否	44.1	25.5	48.5	62.4

5 结论

为了提升 Transformer 在目标检测应用中的性能，以扩大 Transformer 在目标检测中的广泛应用。本文首先针对大尺度特征图难以作为 Transformer 输入的问题，提出层间残差注意力方法，用于实现小尺度特征信息的增强，提高了 DETR 对于目标尤其小目标的检测准确率；接着，对于 Transformer 的输出进行推理的过程中，提出可实现实体挖掘的胶囊推理模型，并利用注意力路由的信息传递方式进行目标类别和位置的预测，最终增强了 DETR 模型的检测性能。此外，为了更好的描述特征和类别标签之间的映射关系，本文提出一种超香肠度量模型，该模型不仅具备较强的非线性映射能力，还能够替换 squansh 激活函数，实现胶囊实体预测的目标存在的概率，进一步提升了模型的检测性能。然而，基于注意力的 transformer 所需要的巨大的计算复杂度，使得其应用具有较大的局限性，同时，类脑启发所构建的超香肠度量模型以一维单纯形作为基本单元，所形成的超曲面具有单一性，未来我们将更加关注于更复杂且有效的超曲面模型的挖掘，探索更高效的注意力实现方法，以实现基于 Transformer 的目标检测在实际中的广泛应用。

参 考 文 献

[1] Zhang Dong-ming, Jin Guo-qin, Dai Feng et. al. Salient object detection based on deep fusion of hand-crafted features. Journal of Computer Science, 2019, 42(9): 2076-2086.
(张冬明, 靳国庆, 代锋, 等. 基于深度融合的显著性目标检测算法. 计算机学报, 2019, 42(9): 2076-2086.)

[2] Viola P, and Jones M. Rapid object detection using a boosted cascade of simple features//Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. Kauai, USA, 2001: 1063-6919.

[3] Girshick R. Fast r-cnn//Proceedings of the IEEE international conference on computer vision. Santiago, Chile, 2015: 1440–1448.

[4] Girshick R, Donahue J, Darrell T, and Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation//Proceedings of the IEEE conference on computer vision and pattern recognition, Columbus, USA, 2014: 580–587.

[5] Redmon J, Divvala S, Girshick R, and Farhadi A. You only look once: unified, real-time object detection//Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, USA , 2016: 779–788.

[6] Ren S, He K, Girshick R, and Sun J. Faster r-cnn: towards real-time object detection with region proposal networks. Advances in neural information processing systems, 2015, 28: 91–99.

[7] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C.-Y, and Berg A. C. Ssd: single shot multibox detector. //Proceedings of the European conference on computer vision, Amsterdam, Netherlands , 2016, 9905: 21– 37.

[8] Kuhn H. W. The Hungarian method for the assignment problem. Naval research logistics quarterly, 1955, 2(1-2): 83–97.

[9] Romera-Paredes B, and Torr P. H. S. Recurrent instance segmentation. //Proceedings of the European conference on computer vision, Amsterdam, Netherlands , 2016, 9910: 312–329.

[10] Stewart R, Andriluka M, and Ng A. Y. End-to-end people detection in crowded scenes//Proceedings of the IEEE conference on computer vision and pattern recognition, Vegas, USA , 2016: 2325–2333.

[11] He K, Zhang X, Ren S, and Sun, J. Deep residual learning for image recognition//Proceedings of the IEEE conference on computer vision and pattern recognition. Vegas, USA , 2016: 770–778.

[12] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A. N, Kaiser Ł, and Polosukhin, I. Attention is all you need. //Proceedings of the 31st Conference on Neural Information Processing Systems, Long Beach, USA, 2017:5998-6008.

[13] Zheng M, Gao P, Zhang R, et al. End-to-end object detection with adaptive clustering transformer. arXiv: 2011.09315, 2020.

[14] Zhu X, Su W, Lu L, et al. Deformable detr: deformable transformers for end-to-end object detection. arXiv: 2010.04159, 2020.

[15] Sun Z, Cao S, Yang Y, et al. Rethinking transformer-based set prediction for object detection//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 3611-3620.

[16] H. S. Seung. Learning continuous attractors in recurrent networks. Advances in neural information processing systems, 1997, 10:654-660.

[17] X Ning, Y Wang, W Tian, et al., A biomimetic covering learning method based on principle of homology continuity. ASP Transactions on Pattern Recognition and Intelligent Systems. 2021, 1(1): 9-16.

[18] Sabour, S.; Frosst, N.; and Hinton, G. E. Dynamic routing between capsules. arXiv preprint arXiv: 1710.09829, 2017.

[19] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, 1943, vol. 5, 115–133.

[20] D.S. Broomhead and D. Lowe. Radial basis functions, multi-variable functional interpolation and adaptive networks. Great malvern, United Kingdom:Royal Signals and Radar Establishment Malvern , 1988 Tech. Rep:4148.

[21] S.-Q. Zhang and Z.-H. Zhou, Flexible transmitter network. Neural Computation, 2021, 33: 2951–2970.

[22]Kisantal M, Wojna Z, Murawski J, et al.

Augmentation for small object detection. arXiv:1902.07296, 2019.

[23]Li J, Liang X, Wei Y, et al. Perceptual generative adversarial networks for small object detection//Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu, USA, 2017: 1222-1230.

[24] Li Z, Peng C, Yu G, et al. Detnet: A backbone network for object detection. arXiv:1804.06215, 2018.

[25] Singh B, Davis L S. An analysis of scale invariance in object detection snip//Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City, USA, 2018: 3578-3587.

[26] Li Y, Li S, Du H, et al. YOLO-ACN: focusing on small target and occluded object detection. IEEE Access, 2020, 8: 227288-227303.

[27] Zhang Y, Bai Y, Ding M, et al. Multi-task generative adversarial network for detecting small objects in the wild. International Journal of Computer Vision. 2020, 128(6): 1810-1828.

[28] Song Z, Zhang Y, Liu Y, et al. MSFYOLO: feature fusion-based detection for small objects. IEEE Latin America Transactions, 2022, 20(5): 823-830.

[29] Ning C, Menglu L, Hao Y, et al. Survey of pedestrian detection with occlusion. Complex & Intelligent Systems, 2021, 7(1): 577-587.

[30] Tian Y, Luo P, Wang X, et al. Deep learning strong parts for pedestrian detection//Proceedings of the IEEE international conference on computer vision. Santiago, Chile, 2015: 1904-1912.

[31] Ouyang W, Zhou H, Li H, et al. Jointly learning deep features, deformable parts, occlusion and classification for pedestrian detection. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(8): 1874-1887.

[32] Wang X, Xiao T, Jiang Y, et al. Repulsion loss: Detecting pedestrians in a crowd//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 7774-7783.

[33] Ke W, Huang D. Improving Object Detection with Inverted Attention//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Snowmass Village, USA, 2020: 1305-1313.

[34] J. Somers. Is ai riding a one-trick pony? Technology review. 2017, 120(6), 29–36.

[35] G. E. Hinton, S. Sabour, and N. Frosst. Matrix capsules with em routing. //Proceedings of the international conference on learning representations workshop. Vancouver, Canada, 2018.

[36] D. Wang and Q. Liu. An optimization view on dynamic routing between capsules. //Proceedings of the international conference on learning representations workshop. Vancouver, Canada, 2018:1-15.

[37] A. Jaiswal, W. AbdAlmageed, Y. Wu, and P. Natarajan. Capsulegan: generative adversarial capsule network//Proceedings of the European conference on computer vision. Munich, Germany, 2018: 526-535.

[38] R. LaLonde and U. Bagci. Capsules for object segmentation. arXiv:1804.04241, 2018.

[39] Lee J, Lee Y, Kim J, et al. Set transformer: a framework for attention-based permutation-invariant neural networks//Proceedings of the International Conference on Machine Learning, Long Beach,USA, 2019: 3744-3753.

[40] Ning, X.; Tian, W.; Li, W.; Lu, Y.; Nie, S.; Sun, L.; and Chen, Z. BDARS capsnet: bi-directional attention routing sausage capsule network. IEEE Access, 2020, 8: 59059–59068.

[41] Ning X, Li W, Li H, et al. Uncorrelated locality preserving discriminant analysis based on bionics. Journal of Computer Research and Development, 2016, 53(11): 2623.
(宁欣, 李卫军, 李浩光, 等. 基于仿生学的不相关局部保持鉴别分析. 计算机研究与发展, 2016, 53(11): 2623.)

[42] Ning X, Li W, Tang B, et al. BULDP: biomimetic uncorrelated locality discriminant projection for feature extraction in face recognition. IEEE Transactions on Image Processing, 2018, 27(5): 2575-2586.

[43] Ning X, Li W, Tian W, et al. Topological higher-order neuron model based on homology-continuity principle//China Symposium on Cognitive Computing and Hybrid Intelligence (CCHI), Xi'an, China, 2019:

161-166.

[44] Calvin W. Handbook of Brain Theory and Neural Networks：Cortical columns, modules, and Hebbian cell assemblies. Cambridge, USA：MIT Press ,1998.

[45] Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollar, P.; and Zitnick, C. L. Microsoft coco: common objects in context. //Proceedings of the European conference on computer vision, Zurich, Switzerland, 2014: 740–755.

[46] Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. End-to-end object detection with transformers. //Proceedings of the European conference on computer vision, Glasgow, USA, 2020: 213–229.

[47] Kingma D P, Ba J. Adam: A method for stochastic optimization. arXiv:1412.6980, 2014.

[48] Loshchilov I, Hutter F. Fixing weight decay regularization in adam//Proceedings of the international conference on learning representations workshop. Vancouver, Canada, 2018.

[49] Tian, Z.; Shen, C.; Chen, H.; and He, T. Fcos: fully convolutional one-stage object detection//Proceedings of the IEEE/CVF international conference on computer vision, Seoul, Korea, 2019: 9627–9636.

[50] Farhadi A, Redmon J. Yolov3: an incremental improvement//Computer Vision and Pattern Recognition. Berlin/Heidelberg, Germany, 2018: 1804-02.

[51] Law, H.; and Deng, J. Cornernet: Detecting objects as paired keypoints//Proceedings of the European conference on computer vision (ECCV). Munich, Germany, 2018: 734–750.

[52] Lin T.-Y, Goyal P, Girshick R, He K and Dollar P. Focal loss for dense object detection//Proceedings of the IEEE international conference on computer vision. Venice, Italy, 2017: 2980–2988.

Ning Xin, Ph.D. ,Associate professor. His current research interests include pattern recognition, computer vision, and image processing.

Tian Weijuan, M.S. ,engineer. Her research interests include machine learning, DNNs, neuron modeling, video and image processing.

Li Weijun, Ph.D., Professor. His research interests include deep modeling, machine art, pattern recognition, artificial neural networks and intelligent system.

Yu Lina, . Ph.D., Assistant professor. Her researches focus on machine learning, deep modeling, and intelligent system.

Background

In recent years, the deep learning has been applied in many computer visions and image processing tasks. Deep learning shows excellent performance, which promotes many researchers to apply it in object detection, including the several popular directions: improve the direction accuracy by updating the network structure, design simple network model based on the transformer, and obtain better detection results through the characteristics of steganalysis. After analyzing the existing transformer-based object detection algorithms, we found that the detection performance for small objects and occluded objects are poor, and the efficiency of most algorithms was low large-scale image processing. To solve the above problems, we analyze the brain-like cognition idea, and proposed a new rain inspired object detection approach with inference pattern, namely utilizing capsule inference and residual augmentation modules (CIRA_DETR).

First of all, the design of inter-layer residual information enhancement allowed us to calculate the discrepancy between different scale of features and enhances the object-related information in small-scale features, which is achieved without increasing the model complexity, thus improving the convergence speed and detection performance of the DETR model. Secondly, in order to be close to the way of human brain thinking, the hierarchy and transmission way of internal knowledge representation in the network is simulated, and a capsule inference module is constructed to combine capsule construction and information routing methods for object category and location inference, which can introduce the relative location information and category association of objects in the process of inference, and thus, can capture global information and enhance the detection performance of targets. Thirdly, based on the brain science inspiration that things

eventually converge to a continuous attractor, a hyper-sausage measure model with stronger nonlinear capability is introduced in the inference process, which is used to form a more flexible and variable hypersurface, so as to describe the mapping relationship between features and labels, and improve the representative ability of model for objects.

This work was supported by the National Nature Science Foundation of China, grant no. 61901436. The project aims to construct view-variant feature extraction method, and solve multi-view image recognition problems. The proposed CIRA_DETR is effective for the detection of small and occluded objects, and the way of mining object entities provides a new thinking way for the viewpoint invariance feature extraction.

联系人：宁欣 15311302805 ningxin@semi.ac.cn