计算机科学
COMPUTER SCIENCE

基于细粒度注意力机制的人与物体交互检测

丁元博, 白琳, 李陶深

引用本文

丁元博, 白琳, 李陶深. 基于细粒度注意力机制的人与物体交互检测[J]. 计算机科学, 2025, 52(11): 141-149.

DING Yuanbo, BAI Lin, LI Taoshen. Human-Object Interaction Detection Based on Fine-grained Attention Mechanism [J]. Computer Science, 2025, 52(11): 141-149.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240900113

基于细粒度注意力机制的人与物体交互检测

丁元博　白　琳　李陶深
广西大学计算机与电子信息学院　南宁 530004
(2603491489@qq.com)

摘　要　细粒度信息作为一种上下文信息,能够辅助模型识别相对空间关系相似的人与物体交互动作。然而,如何利用这一关键线索统一建模多尺度特征图上不同粒度的特征信息,仍然是人与物体交互检测精度进一步提升面临的主要挑战之一。为了解决这一问题,提出了一种基于细粒度注意力机制的人与物体交互检测模型(FGDHOI)。该模型在细粒度信息的指导下强化局部特征,融合不同尺度的特征图,通过可变形注意力机制自动学习图像内容,并建模不同粒度特征之间的长距离依赖关系,从本质上提高了人与物体交互检测模型的精度。在 V-COCO 和 HICO 数据集上进行了广泛的定性、定量及消融实验。实验结果表明,所提出的方法相比基准模型,在 V-COCO 数据集上 mAP 提升了 7.7 个百分点,在 HICO 数据集 3 项指标上 mAP 分别提升了 7.43 个百分点、7.5 个百分点和 7.85 个百分点。

关键词:深度学习;人与物体交互检测;细粒度信息;注意力机制
中图分类号　TP391

Human-Object Interaction Detection Based on Fine-grained Attention Mechanism

DING Yuanbo,BAI Lin and LI Taoshen
School of Computer and Electronic Information,Guangxi University,Nanning 530004,China

Abstract　Fine-grained information,as a kind of contextual information,can assist models in recognizing human-object interactions with similar relative spatial relationships. However,how to utilize this key cue to uniformly model feature information of different granularities on multi-scale feature maps remains a critical challenge that hinder further improvement of human-object interaction detection accuracy. To address this problem,this paper proposes a human-object interaction detection model based on fine-grained attention mechanism. The model strengthens local features under the guidance of fine-grained information. It fuses feature maps of different scales and automatically learns image content through a deformable attention mechanism. Additionally,it models the long-range dependencies between features of various granularities,essentially improving the accuracy of the human-object interaction detection model. Extensive experiments are conducted on the V-COCO and HICO datasets. The experimental results show that the proposed method has increased the mAP by 7.7 percentage points on the V-COCO dataset,and the mAP has increased by 7.43,7.5 and 7.85 percentage points on the HICO dataset compared to the baseline models.

Keywords　Deep learning,Human-Object interaction detection,Fine-grained information,Attention mechanism

1　引言

人与物体交互关系(Human-Object Interaction,HOI)检测是图像内容理解任务的重要子任务之一,其在运动员辅助训练、智能监控、信息检索等方面有着巨大的应用价值。人与物体交互检测任务的目标是检测出图像中〈人,物体,交互关系〉的三元组[1-2]。由于人与物交互关系的多样性,单纯依靠外观特征和实例级空间关系,往往难以区分相似多变的交互关系,HOI 任务仍然具有很大的挑战性。

人与物体交互任务需要模型对图像内容有更深刻的理解。为了实现这一目标,近年来研究人员发现人体姿态信息、人与物体空间相对位置信息等细粒度信息有利于模型对图像的理解。Wan 等[3]提出了多层关系网络(Pose-aware Multi-level Feature Network,PMFNet),利用人体姿态信息来捕获全局空间关系配置,以辅助 HOI 检测任务。Li 等[4]提出了一种交互识别方法(Transferable Interactiveness Knowledge Network,TIN),其核心思想是利用交互网络从多个 HOI 数据集学习一般的交互知识。交互网络使用人、物体相对位置信息、人体姿态信息辅助交互关系检测。

上述研究[3-4]证明了人体姿态信息的加入对人与物体交

到稿日期:2024-09-18　返修日期:2024-12-02
基金项目:国家自然科学基金(61966003)
This work was supported by the National Natural Science Foundation of China(61966003).
通信作者:白琳(bailin@gxu.edu.cn)

互关系检测任务的有效性。但是上述方法在对包括人体姿态信息在内的细粒度信息的利用还存在以下问题。

1)对细粒度信息的使用依赖于人体、物体外观信息和空间相对位置关系。没有利用人体关键点信息、物体细粒度信息,以及两者之间的空间关系信息,导致难以区分外观信息和相对空间位置信息相似的交互关系。如图1所示,在图中有两组人与物体交互实例,它们具有相似的人体、物体实例外观信息和空间位置信息。左边一组的交互关系为坐在摩托车上,右边一组为站在摩托车旁。在这种情况下引入细粒度信息辅助模型识别人与物体交互关系是十分关键的。

图1 外观特征相似的人与物体交互实例

Fig. 1 Instances of human-object interaction with similar appearance features

2)无法对不同粒度特征进行统一建模。基于特征裁剪的多阶段检测方法,是目前利用细粒度信息的人与物体交互检测的最主流方法。这类方法在第一阶段利用目标检测器生成具有相应置信度的人/对象/细粒度信息边界框。在第二阶段,通过解析裁剪特征来检测人与物体之间的交互。然而,这类方法存在的问题是,细粒度信息、空间和上下文信息与全局特征之间的交互被忽略,造成同一语义下的特征被分开,从而降低了模型精度。

针对以上问题,本文提出了基于细粒度注意力机制的人与物体交互检测(Fine-grained Deformable Attention Mechanism for Human-Object Interaction Detection, FGDHOI)模型,如图2所示。首先,通过本研究团队前期研究成果SKD-Pose[5]模型获得包括人体姿态信息和物体中心点等细粒度信息,以及人体边界框和物体边界框。接着,设计了一个细粒度特征局部增强模块,在细粒度信息指导下强化高分辨率特征图中的局部特征。然后,通过基于注意力机制的特征融合,融合多尺度特征图。最后,使用可变形注意力机制在细粒度信息的指导下,在融合特征图上进行特征提取和全局关系建模,通过多层感知机制推断交互关系。

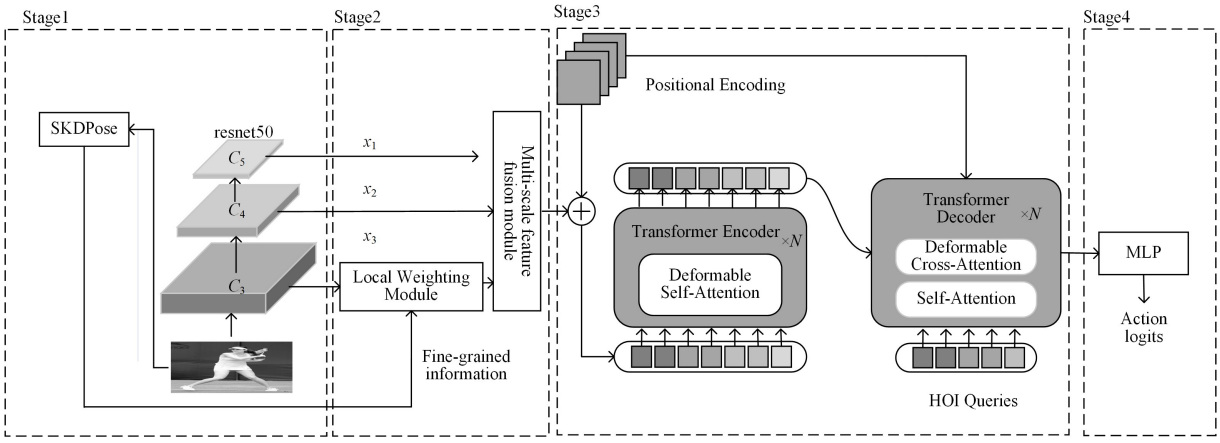


图2 基于细粒度信息的人与物体交互检测网络

Fig. 2 Human-object interaction detection network based on fine-grained information

本文的主要贡献如下:

1)针对人与物体交互检测任务,提出了一种细粒度特征局部增强的细粒度信息使用方法。该模型能够对不同粒度的特征信息进行统一建模。

2)提出了基于注意力机制的特征融合模块,能在全局和局部上下文信息的指导下进行特征融合,解决了多尺度特征融合时上下文信息聚合偏向性的问题。

3)使用可变形注意力机制,建模多尺度特征图不同粒度特征之间的依赖关系,生成上下文信息丰富的视觉特征。实验表明,该视觉特征能够提升人与物体交互检测的精度。

2 相关工作

2.1 人与物体交互

早期的HOI检测研究主要依赖于手工制作的局部特征,如颜色、方向梯度直方图[6](Histogram of Oriented Gradient, HOG)、尺度不变特征变换算法[7](Scale Invariant Feature Transform, SIFT),然而这类特征训练出的模型泛化能力极差。Gupta等[1]首次提出了专门用于HOI的标准数据集。随着深度学习的快速发展,研究人员可以利用神经网络从大规模的数据集中自动提取特征,推动HOI检测任务进入新的阶段。多流结构是基于深度学习HOI检测任务框架中非常有代表性的一种方法,该方法对视觉特征进行多分支的特征提取与建模,再融合各分支的结果。Gao等[8]提出了ICAN———一种以实例为中心的注意力机制网络,包含3个分支:人分支、物分支和成对关系分支。在以各自实例为中心的注意力机制的指导下,人和物两个分支分别推理交互得分。成对关系分支通过掩码技术得到人-物的空间配置,融合人的外观特征推理交互得分。最后融合3个分支,推理出交互关系。Gkioxari等[9]提出了InteractNet模型,该模型主要基于人-物的外观信息来推理交互关系。该模型以Faster-RCNN[10]为基础,第一个分支得到图像中的物体特征和人物特征,第二个分支利用人物外观特征推断交互行为的种类,第三

个分支结合物体特征和人物特征的相对空间关系辅助第二支路推断交互行为。

然而，这些模型在检测人与物体之间的交互关系时，完全依赖于人体和物体的外观信息，导致在检测外观相似的交互关系时表现较差。为了解决上述问题，Wan 等[3]提出了姿态维度多级特征网络 PMFNet，该网络有两个主要模块：整体模块和放大模块。整体模块包含 4 条支路：人体支路、物体支路、联和支路以及由人体姿态信息指导获取全局空间配置的空间支路。放大模块在人体姿态线索的指导下，通过注意力机制提取局部特征。两个模块将不同维度、不同语义的特征融合到一起，生成更加多维度和更有鲁棒性的推理结果。Li 等[11]提出了 HGNN 模型，从局部到整体显式地对人体关键区域以及人和物构成的场景图进行建模，利用注意力机制筛选人体关键点，来增强局部特征。Lin 等[12]提出了一种行为感知的注意力机制，来更好地区分细粒度行为之间的差异。Sun 等[13]提出了融合空间-语义知识的多级调节网络，通过融合多模态特征来识别交互关系。

但是，这些方法通常将细粒度特征提取与全局特征提取的过程分开，导致图像上下文信息在特征提取过程中可能会丢失，也使得细粒度特征和全局特征无法统一建模。

为了探究不同人物对之间的联系，一些研究提出了使用图的概念来处理 HOI 检测任务。Qi 等[14]提出了 GPNN 网络，首次将图网络引入 HOI 检测任务。该网络将人体、物体实例定义为图节点，用节点之间的边编码交互关系，通过连接函数学习最优的图结构，从而推理出交互关系。Wang 等[15]在图结构中加入了同构实体和异构实体的考量，使用连接异构节点的边编码人-物之间的交互关系，同构节点则编码人与人、物体与物体等同类关系，以获得更充分的上下文信息。Ulutan 等[16]提出了视觉空间图网络，该网络有 3 个分支，其中视觉分支和空间分支的结果合并可以得到高质量的视觉特征，基于图的分支建模人-物对的交互关系，该图分支只对视觉差异大的实例之间建模。这些方法通过构建优质的图模型，有效地辅助了交互关系的推理。然而，图模型无法精确表示人体与物体交互时的不同作用，同时也忽略了背景信息和细粒度信息等图像上下文信息。

为了解决以上问题，本文研究了图像上下文特征及更有效的关系建模方法。针对现有研究中的不足，提出了一种新方法，通过局部增强细粒度特征，并采用可变形注意力机制自动提取高级视觉特征。实验结果表明，包含丰富上下文信息的视觉特征显著提高了 HOI 检测的精度。

2.2 Transformer 模型

多头自注意力模型 Transformer[17]在自然语言处理领域取得了重大成功。注意力机制具有强大的长距离依赖建模能力，其在 Transformer 中反复应用。在自注意力机制层中，首先对输入向量进行线性变换得到 3 个不同的向量：查询向量 query、关键字向量 key、值向量 value。自注意力机制的计算式如下：

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{SoftMax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V} \tag{1}$$

其中，$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 分别是由 query，key，value 组成的矩阵，$\sqrt{d}$ 是缩放因子，d 的大小与 $\boldsymbol{K}$ 矩阵的维度一样。通过 $\boldsymbol{Q}$ 矩阵和 $\boldsymbol{K}$ 矩阵的转置相乘除以缩放因子得到注意力分数矩阵，注意力分数矩阵通过归一化操作得到注意力权重矩阵，这些权重代表其他序列与当前序列的相关性。多头注意力机制在自注意力机制的基础上将模型划分为多个头，形成多个子空间，使模型可以关注不同信息。多头注意力机制公式如下：

$$\boldsymbol{Q}_i=\boldsymbol{Q}\boldsymbol{W}_i^{Q},\boldsymbol{K}_i=\boldsymbol{K}\boldsymbol{W}_i^{K},\boldsymbol{V}_i=\boldsymbol{V}\boldsymbol{W}_i^{v},i=1,\cdots,8 \tag{2}$$

$$head_i=Attention(\boldsymbol{Q}_i,\boldsymbol{K}_i,\boldsymbol{V}_i) \tag{3}$$

$$MultiHead(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=Concat(head_1,\cdots,head_8)\boldsymbol{W}^{O} \tag{4}$$

近年来，不少研究将 Transformer 模块引入了 HOI 检测任务。Zou 等[18]将卷积神经网络和编码器作为一个整体的特征提取器，用于提取图像特征并编码全局信息。解码器为所有目标查询解码出一个视觉特征，多层感知机根据这些特征检测出〈人，物体，交互关系〉的三元组。Tamura 等[19]提出了 QPIC 网络架构，通过成对的查询信息检测人与物体交互关系，提高了模型在困难场景下的关系建模能力，达到了最优解的效果。

这两种方法的成功主要来自于 Transformer 强大的建模能力，但是它们没有考虑使用人体姿态信息这一对人-物交互图像内容理解十分有效的线索。同时，Transformer 的计算复杂度与输入图像的尺度成二次相关，所以输入只能是主干神经网络中的高级视觉特征。而经过反复下采样的高级视觉特征中，细粒度特征丢失严重。

为了克服这些问题，本文增强了 Transformer 输入特征中的细粒度局部特征，并通过可变形注意力机制来学习不同粒度特征之间的联系，使得 Transformer 能够建模人体部位与物体之间的细粒度空间信息，得到上下文信息更加丰富的特征。

3 本文方法

本文将一张图片中的 HOI 实例定义为一个五元组〈***human class***，***object class***，***action class***，***human box***，***object box***〉。本文提出的 FGDHOI 模型的整体架构如图 2 所示。FGDHOI 模型一共包含 4 个部分。

第一阶段为特征提取阶段。对于给定图像 I，提取多尺度特征图和细粒度信息作为下一部分的输入。本文使用 resnet50 作为主干网络（Backbone），从 resnet50 的 C_3-C_5 阶段提取到多尺度特征图 $\boldsymbol{F}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3\},\boldsymbol{x}_i\in\boldsymbol{R}^{C\times H_i\times W_i}$。同时，将图片输入 SKDPose 模型[6]，得到细粒度信息。

第二部分为细粒度信息增强与多尺度特征融合阶段，该部分中的细粒度特征局部增强模块和基于注意力机制的特征融合模块为本文模型的核心模块。本文模型选择包含丰富局部特征的底层特征 $\boldsymbol{x}_3$ 输入特征局部增强模块，在细粒度信息和缩放因子的指导下对 $\boldsymbol{x}_3$ 中的细粒度特征进行增强，得到 $\boldsymbol{x}_3'$。图像任务中特征尺度的不同对任务的检测结果有显著影响。高层特征包含语义信息，底层特征则包含大量上下文信息，两者都是人与物体交互检测任务的重要线索。故在特征融合模块中，将 $\boldsymbol{x}_3'$，$\boldsymbol{x}_1$，$\boldsymbol{x}_2$ 特征融合得到的特征作为下一部分的输入。

第三部分为交互关系高级视觉特征生成部分。在该部

分，本文基于 Transformer 设计了交互关系高级视觉特征生成器。交互关系高级视觉特征生成器包含编码器和解码器两个模块。编码器从输入中提取出全局特征，解码器通过可学习的 HOI Queries 在全局特征中解码出高级视觉特征。

第四部分为交互关系预测阶段。本文基于高级视觉特征，通过多层感知机预测交互关系动作的置信度。通过交叉熵损失函数计算动作类置信度和真实标签之间的损失。

本章将分别介绍 FGDHOI 模型的 3 个主要模块：细粒度特征局部增强模块、基于注意力机制的特征融合模块、交互关系高级视觉特征生成器。

3.1 细粒度特征局部增强模块

目前主流的基于多阶段特征裁剪的方法，不仅不能建模不同粒度特征之间的联系，还会导致图像中的上下文信息丢失[31]。为此，本文提出了细粒度特征局部增强模块，在全局特征图中突出细粒度特征。在后文中，通过可变形注意力机制自动关注细粒度特征。

在本文中，SKDPose 检测出的细粒度信息被作为交互关系检测的先验知识。人框和物框可以用 $\boldsymbol{H}=\{\boldsymbol{h}_1,\cdots,\boldsymbol{h}_n\}$，$\boldsymbol{O}=\{\boldsymbol{o}_1,\cdots,\boldsymbol{o}_m\}$两组序列来表示，其中 n 表示图像中人体数量，m 表示图像中物体数量。得到的细粒度信息包括人体关键点坐标和物体中心点坐标。图像中第 i 个人物的关键点序列可以表示为$\boldsymbol{J}_{h_i}^{K}=\{\boldsymbol{j}_{h_i}^{k}\mid k=1,\cdots,a\}$，其中 a 表示人体关键点数量，第 e 个物体的中心点可以表示为g^{o_e}。

本文选择包含细粒度特征最丰富的底层特征 $\boldsymbol{x}_3$ 作为加权模块的输入。首先，生成一张大小与$\boldsymbol{x}_3$一样的张量热图(heat_map)，再通过缩放因子找到关键点在热图上的位置。根据人框 h、物框 o 的大小来决定对应人体关键点、物体中心点的增强区域大小，以关键点坐标为中心设置高斯权重得到热图，将热图加到$\boldsymbol{x}_3$上完成增强，得到增强后的特征图$\boldsymbol{x}_3{}'$。

在细粒度特征局部增强模块中，$\boldsymbol{x}_3$ 的对应热图上对于每一个物体中心点和人体关键点$\boldsymbol{k}^{o_e}$和$\boldsymbol{j}_{h_i}^{K}$都需要生成高斯分布区域。高斯分布区域生成如式(5)、式(6)所示：

$$d_i=\sqrt{(x-x_i)^2+(y-y_i)^2} \tag{5}$$

$$w_i=\mathrm{e}^{-\frac{d_i^2}{2\sigma^2}} \tag{6}$$

其中，d_i代表中心点到任意像素点的距离；w_i代表任意点像素点的权值；σ是控制高斯分布的参数，在本文中 σ 由增强区域的大小决定。细粒度特征局部增强模块的算法流程如算法 1 所示。

算法 1 细粒度特征局部增强算法

输入：人框集合 $\mathbf{H}=\{\mathbf{h}_1,\cdots,\mathbf{h}_n\}$、物框集合 $\mathbf{O}=\{\mathbf{o}_1,\cdots,\mathbf{o}_m\}$、人体部位集合$\mathbf{J}^{K}_{h_i}=\{\mathbf{j}^{k}_{h_i}\mid k=1,\cdots,a\}$、物体中心点$g^{o_e}$、特征图$\mathbf{x}_3$

输出：特征图$\mathbf{x}_3{}'$

1. /* 一张特征图的局部增强 */
2. 初始化：生成一张大小与$\mathbf{x}_3$一样的张量热图 heat_map
for $h_i\in\mathbf{H}$,
3. part_size=$h_i/20$;
4. for $\mathbf{j}^{k}_{h_i}\in\mathbf{J}^{K}_{h_i}$:
5. 根据式(6)在 heat_map 上生成关键点的局部高斯区域;
6. end for
7. end for
8. for $o_g\in O$
9. part_size=$o_g/20$;
10. 根据式(6)在 heat_map 上生成关键点的局部高斯区域;
11. end for
12. 将特征图x_3与 heat_map 广播乘法的值与x_3相加，得到特征图$\mathbf{x}_3{}'$。
13. return $\mathbf{x}_3{}'$

3.2 基于注意力机制的特征融合模块

目前，特征金字塔网络(Feature Pyramid Network，FPN)及其相关工作是特征融合的主流方法。然而，这些工作的研究集中在通过不同的路径融合不同层次的特征，其中，融合操作是基于简单的线性组合。处理多尺度特征融合时，相比线性组合的方法，本文提出的特征融合方法能够根据特征生成权重，指导特征进行有偏向性的融合。如图 3 所示，本文设计了局部支路和全局支路两条支路来生成权重，使得特征融合模块可以兼顾局部特征信息和全局通道信息。

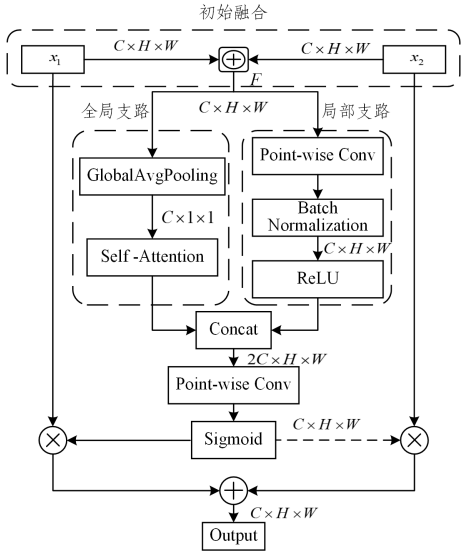


图 3 特征融合模块

Fig. 3 Feature fusion module

在全局支路中同时提取所有通道上的信息，在局部支路中分别对每一通道上的特征图进行特征提取。以$\boldsymbol{x}_1$，$\boldsymbol{x}_2$特征融合为例，如图 3 所示，首先将$\boldsymbol{x}_1$与$\boldsymbol{x}_2$初始融合得到初始融合特征 F，采用逐元素相加的方法进行初始融合。接着，将 F 分别送入通道全局支路和局部特征支路。全局支路的目的是显式地建模通道之间的依赖性，从而得到不同通道之间的上下文信息。全局支路中使用一个全局平均池化，将包含全局信息的特征图压缩到一个 $C\times 1\times 1$ 的特征向量 $\boldsymbol{L}$ 上。本文在全局支路上使用长距离依赖建模能力更强的自注意力机制代替卷积操作，自注意力机制如式(1)所示。在局部支路中，通过逐点卷积(Point-wise Conv)来聚合局部的上下文信息。上述两条支路中的计算式如下：

$$L(F)=\delta(\beta(PWConv_1(F))) \tag{7}$$

$$G(F)=Attention(GAP(F)) \tag{8}$$

其中，F 代表x_1和x_2按像素求和的初始融合特征，$L(F)$代表局部特征信息，$G(F)$代表通道上的全局信息，β 表示 Batch

Normalization，δ 表示 ReLU 激活函数，GAP 是全局平均池化操作，$PWConv_1$ 的大小为 $C\times C\times 1\times 1$。在得到 $L(F)$ 和 $G(F)$ 之后，将两者拼接到一起，再通过逐点卷积将其降维到 $C\times H\times W$。通过 Sigmoid 操作得到融合权重：

$$M(F)=\phi(PWConv_2(Concat(L(F),G(F)))) \tag{9}$$

其中，$M(F)$ 代表融合权重，ϕ 代表 Sigmoid 操作，$PWConv_2$ 的大小为 $2C\times C\times 1\times 1$。最终的融合过程如式(10)所示：

$$Z=M(F)\otimes x_1+(1-M(F))\otimes x_2 \tag{10}$$

其中，Z 代表最终的融合特征，$1-M(F)$ 在图 3 中用虚线表示，$\otimes$ 代表广播乘法。本文在特征融合时使用的权重兼顾了全局通道信息以及局部特征信息，这有助于模块关注背景信息相对简单的细粒度信息。

3.3 交互关系视觉特征生成器

本文的交互关系特征生成器基于 Transformer 模型，为编码器-解码器结构。编码器接收融合特征作为输入，利用可变形注意力机显式地建模图像特征之间的关系，生成全局特征。在图 2 中，解码器将 N 个可学习的位置向量作为 HOI 查询(HOI query)，并将其转换为 N 个输出向量。在传统的注意力机制中，权重是通过对固定位置特征进行注意力计算得到的。而在可变形注意力中，可以根据输入特征动态地调整注意力模型的形状和大小。相比自注意力机，可变形注意力机制有以下两点优势。首先，可变形注意力机制用局部采样代替了全局采样，交互关系视觉特征生成器通过学习可以更多地关注细粒度局部特征。其次，可变形注意力机制降低了 Transformer 的计算成本，在相同算力下可以输入更丰富的特征信息到交互关系视觉特征生成器中。在编码器中，可变形自注意力机制的 Q 为所有特征图上的像素，V 是根据 Q 得到的采样值，其计算式如下：

$$\Delta p=\boldsymbol{Q}\boldsymbol{W}^{p} \tag{11}$$

$$V=Samp(Z,p+\Delta p)\boldsymbol{W}^{v} \tag{12}$$

$$DeformAttn(Q,V)=\mathrm{softmax}(\boldsymbol{Q}\boldsymbol{W}^{A})V \tag{13}$$

其中，参数矩阵 $\boldsymbol{W}^{p}$，$\boldsymbol{W}^{V}$，$\boldsymbol{W}^{A}$ 都可以理解为线性投影。对于 Q 上任一点，设其为参考点 p，有 $p\in R^{(N_H+N_L)\times N\times 2}$。参考点 p 加上由 Q 线性投影得到的偏移量 Δp，可以得到最终的采样点。$Samp$ 是指通过双线性插值的方法在采样点周围对全局特征图 Z 进行采样。与式(1)相比，可变形的注意力机制的注意力权重由 Q 线性投影直接得到，不需要 K 的参与。

解码器结构如图 4 所示。解码器主要由 3 个部分组成：自注意力层、交叉注意力层和前馈神经网络层。每一层都采用残差结构。自注意力层表示如式(1)所示，自注意力层可以有效建模特征间的长距离依赖。可变形交叉注意力机制将编码器的输出 S 作为全局特征。根据偏移量和采样点对特征进行采样，采样值乘注意力得到最后的输出。可变形交叉注意力机制的计算式如式(13)所示，与可变形自注意力机制不同的是，其将可学习的向量作为 Q，编码器的输出 S 作为 V。前馈神经网络层由两个线性层组成，线性层之间通过 ReLU 激活函数连接，以增加非线性。前馈神经网络层能够对特征中的每个位置进行独立的变换和抽象，从而更好地捕捉局部特征和语义信息。

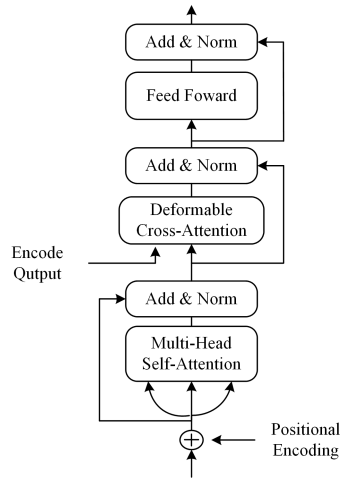


图 4 解码器结构

Fig. 4 Structure of decoder

4 实验过程

本章将详细介绍人与物体交互检测的具体实验操作细节与实验结果。首先，介绍实现设置、数据集、实现细节；然后，通过在 V-COCO 数据集和 HICO 数据集上与先进的方法进行比较，来定量评估本文模型；随后，通过注意力权重可视化和几个定性结果来证明本文方法的有效性；最后，进行消融实验，评估模型中各个组件的效果。

4.1 实验软硬件环境和模型参数设计

本文实验环境如表 1 所列。

表 1 实验环境

Table 1 Experimental environment

实验环境	配置信息
操作系统	Ubutun16.04.6 LTS
GPU 显卡	Tesla V100_SXM2
CUDA	CUDA Version:10.2
Python	Python 3.7
深度学习框架	PyTorch 1.8
开发工具	PyCharm 2021.3.1

本文算法模型使用开源深度学习开发框架 PyTorch 实现，在两张 Tesla V100 图形处理器(GPU)上训练，单张显卡显存为 16GB。可变形注意力机制的头数和采样偏移量本文依照 Deformable DETR 的设计设置为 $M=8$，$K=4$。编码器层数设置为 6，查询嵌入设置为 300。使用 AdamW 优化器，初始学习率为 2×10^{-4}，权值衰减为 1×10^{-4}。对于 HICO-DET 数据集，本文训练 200 个 epochs，在第 110 个 epochs 衰减学习率。对于 V-COCO 数据集，本文训练 150 个 epoch，在第 60 和第 120 个 epoch 衰减学习率。

4.2 数据集与评价指标

1)数据集

本文在 HICO_DET 和 V-COCO 两个数据集上进行了实验，以评估本文模型的效果。HICO_DET 数据集包含 47776 张图片，其中 38118 张图片作为训练集，9658 张图片作为测试集。该数据集有 600 个 HOI 交互类别，117 个动作类别和 80 个对象以及超过 15 万个人物对。在 600 个 HOI 类别中有 138 个稀有类。V-COCO 数据集是 COCO 数据集的子集，该

数据集从 COCO 数据集中挑选出了 10000 张以上的图片，其中训练集包含 5400 图片，测试集包含 4 946 张图片。该数据集为每个人物编码了 29 个动作的 0-1 标签，其中有 5 种是独立动作，不与物体交互。

2)评价指标

在 HICO-DET 数据集中有两种不同的评价指标，默认(Default)和已知对象(Know Object)。默认对象评价指标考虑图像中出现的所有对象的 AP，已知对象则考虑已经在数据集中包含的 80 个对象类别和人物对象。本文使用对象平均精度(mAP)来评价两个数据集的模型性能。对于 V-COCO 数据集，当且仅当 HOI 检测准确地定位了人和物体(即预测框与真实标签之间的交互-联合(IOU)比大于 0.5)并正确地预测了相互作用时，才认为它是真阳性的。

4.3 实验结果分析

为了评估本文框架在 HOI 检测性能上的表现，本节将对本文框架与当前先进方法进行定量比较，并报告在 V-COCO 和 HICO-DET 数据集上的 mAP 评分。除了定量结果外，还将提供数据集上的一些定性示例。

4.3.1 V-COCO 数据集实验结果

在 V-COCO 数据集上的实验结果如表 2 所列，本文提出的框架在交互动作检测中 mAP 达到了 60.6%，相比其他方法提升了 1.8 个百分点，表现优异。

与本文的基准模型 HOI-Trans 相比，本文模型 mAP 提升了 7.7 个百分点，进一步验证了本文方法的有效性。在表 2 中可以明显看出，基于细粒度信息的模型架构，如 PMFNet，PFNet 和 MLCNet，在与先进模型的比较中取得了显著的成绩，证明了细粒度信息对提升人与物体交互关系检测精度的重要性。然而，与这些模型相比，本文模型仍然具有显著的精度优势。这可以归因于几个因素：首先，本文模型充分利用了细粒度信息，不仅关注人体关键点与物体中心点之间的关系，还考虑了人体关键点内部的关系；其次，本文采用了局部增强细粒度特征和可变形注意力机制，自动学习关键局部与全局特征之间的长距离依赖，实现了对不同粒度特征的统一建模；最后，与当前先进的基于 Transformer 的 HOI 模型 QPIC 相比，本文模型 mAP 仍然提升了 1.8 个百分点，进一步显示了本文模型在全局建模能力上的优势。

表 2 基于 VCOCO 数据集的定量实验

Table 2 Quantitative experiments based on VCOCO dataset

Model	Backbone	mAP_{role}/%
InteractNet[10]	ResNet-50-FPN	40.0
ICAN[9]	ResNet-50	45.3
TIN[5]	ResNet-50	47.8
HGNN[11]	ResNet-50	50.9
PMFNet[4]	ResNet-50-FPN	52.0
RPNN[21]	ResNet-50-FPN	47.5
PFNet[22]	ResNet-50	52.8
MLCNet[13]	ResNet-50-FPN	55.2
QPIC[19]	ResNet-50	58.8
ViPOL[24]	ResNet-50	57.4
FGAHOI[25]	ResNet-50	59.0
HOI-Trans[18]	ResNet-50	52.9
FGDHOI(Ours)	ResNet-50	60.6

4.3.2 HICO 数据集实验结果

表 3 列出了本文方法在 HICO-DET 数据集上与先进方法的定量比较。

表 3 基于 HICO 数据集的定性实验

Table 3 Qualitative experiments based on HICO dataset

(%)

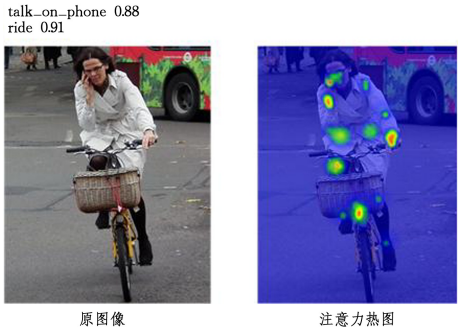
Architecture	Method	Backbone	Default		
			Full	Rare	Non-Rare
Fine grained constraints	TIN[5]	ResNet-50	17.22	13.51	18.32
	RPNN[20]	ResNet-50	17.35	12.78	18.71
	PMFNet[4]	ResNet-50-FPN	17.46	15.65	18.00
	PFNet[21]	ResNet-50	20.05	16.66	21.07
	MLCNet[13]	ResNet-50-FPN	17.95	16.62	18.35
Transformer-Based	HOTR[23]	ResNet-50	25.10	17.34	27.42
	AS-Net[22]	ResNet-50	28.87	24.25	30.25
	QPIC[19]	ResNet-50	29.07	21.85	31.23
	FGAHOI[25]	ResNet-50	29.94	22.24	32.24
	EoID[26]	ResNet-50	28.91	26.66	29.27
	HOI-Trans[18]	ResNet-50	23.46	16.91	25.41
FGDHOI(Ours)		ResNet-50	30.89	24.41	33.26

本文模型的 mAP 分别为 30.89%，23.47%，33.85%，分别以 7.43 个百分点、7.5 个百分点、7.85 个百分点的涨幅高于基准模型 HOI-Trans。如表 3 所列，在与基于细粒度特征的 HOI 模型的比较中，本文模型均有超 10 个百分点的涨幅。推测原因是，上述基于细粒度信息的 HOI 检测模型，对细粒度信息之间的建模依赖于特定的成对人体部位之间的关联，往往忽略了细粒度信息与全局背景知识之间的关联，而本文提出局部加权配合注意力机制自动提取视觉组合的方案兼顾了细粒度特征内部之间的关联以及与全局背景信息的关联。在与基于 Transformer 的 QPIC 检测的比较中，本文模型以 1.82 个百分点、2.56 个百分点和 2.03 个百分点高于基准模型。Transformer 的计算成本与输入特征的尺寸成二次相关，现有的先进方法将特征金字塔顶层尺度最小的特征输入 Transformer，会导致细粒度特征丢失。而本文通过特征融合和可变形注意力机制结合的方法解决了该问题。此外，多模态大模型研究的突破给了人与物体交互领域的研究人员新的启发。其中，EoID[26] 使用预训练的多模态模型 CLIP[27] 提取视觉和文本特征，使用 Transformer 推理交互关系，在解决罕见类交互关系识别问题和零样本学习两方面取得了不错的进展。EoID 模型在罕见类 HOI 关系这一指标上 mAP 相比本

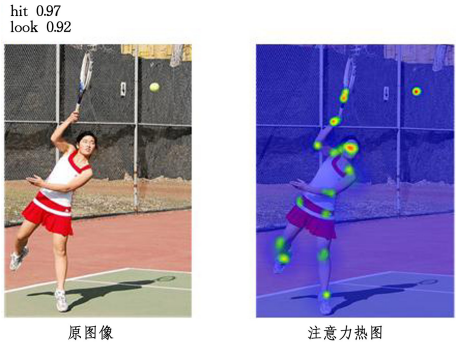
文模型 FGDHOI 高出 2.25 个百分点，但在其他两项指标上分别低 1.98 个百分点和 3.99 个百分点。分析原因是，文本特征的引入确实能够帮助解决训练样本的“长尾问题”，但是也引入了一些杂糅特征。本文方法能够最大程度地减少主干网络到交互关系特征推理阶段之间特征传递的损失，所以在整体精度上优于 EoID 模型。

4.3.3 定性实验

为了验证 FGDHOI 的“细粒度特征增强＋可变形注意力机制”方案的有效性，对其进行可视化实验。图 5 列举了两组交互动作实例，在注意力热图中突出的斑块代表注意力分数，颜色注意力分数越高代表该区域特征对交互关系的检测重要性越强。图 5(a)中注意力机制对于手部和物体给予了更高的注意力分数，有助于 talk_on_phone 和 ride 两种交互关系的检测。在图 5(b)中，hit 交互动作中的球拍和手部获得了较高的注意力分数，对于 look 交互动作，注意力机制对脸部和网球给予更多关注。由此可见，“细粒度特征增强＋可变形注意力机制”方案，有助于模型关注到对交互关系有重要影响的局部特征，特别是在一张图包含多组交换关系的复杂情况下，能够辅助模型更加精确地关注到发生交互关系的人体部位和物体，提高交互动作分类分数。此外，本文还提供两组可视化对比实验结果，其中绿色框表示人体边界框，蓝色框为物体边界框，如图 6 所示。在图 6(a)中，列举了 3 张 FGDHOI 能够正确识别交互关系而基准模型识别错误的图片。在图 6(b)中，图片背景复杂，基准模型对人-物交互对做出了错误检测，即使输出了正确的交互关系，但是没有正确识别出发生交互的人体或物体。而在 FGDHOI 模型中，基于细粒度信息的指导大大降低了发生该错误的可能。



(a)骑自行车交互动作图



(b)打网球交互动作图

图 5 可视化实验结果

Fig. 5 Visualization of the experimental results

图 6 V-COCO 验证集部分人物交互检测结果(电子版为彩图)

Fig. 6 V-COCO validation set of human-object interaction detection results

4.3.4 消融实验

为了研究细粒度信息、特征融合模块，以及可变形注意力机制对本文框架的影响，本小节在 VCOCO 数据集上进行消融实验。本文模型是以 HOI-Trans 为基础进行改进的，所以选择 HOI-Trans 作为消融实验的基础模型。消融实验结果如表 4 所列。

表 4 FGDHOI 组件消融实验

Table 4 Ablation experiment of FGDHOI component

模型	细粒度特征局部强化	可变形注意力机制	特征融合模块	*mAP*/%
HOITrans				52.9
FGDHOI_01	√			54.8
FGDHOI_02		√		51.2
FGDHOI_03	√	√		57.9
FGDHOI	√	√	√	60.6

FGDHOI_01 的全局特征为主干神经网络 resnet50 第五阶段特征图，使用 1×1 卷积将其通道数从 2048 减少到 256，展平成序列输入 Transformer Encoder。FGDHOI_01 相比 HOI-Trans mAP 提升了 1.9 个百分点，证明了本文提出的细粒度特征局部强化方法能够增强模型对细粒度特征的关注，从而使模型更精准地检测人与物体交互关系。FGDHOI_02 将 HOI-Trans 中的注意力机制替换为可变形注意力机制，mAP 下降了 1.7 个百分点。推测精度下降的原因为：在没有其他信息指导下的可变形注意力机制无法准确采样到有效的局部信息，难以建模不同粒度特征之间的依赖。FGDHOI 在 resnet50 第三阶段的输出特征上增强局部特征，增强后与

resnet50 第四、第五阶段的特征图融合。通过 FGDHOI_03 与 FGDHOI 之间对照可以看出，本文的特征融合模块带来了 2.7 个百分点的 mAP 增长，证明了多尺度特征图能为人与物体交互检测提供更丰富的线索。

结束语 本文提出了一种基于细粒度注意力机制的人与物体交互检测方法，旨在充分利用细粒度信息指导模型对图像进行更精确的交互关系检测。本文方法包括以下主要模块：1)细粒度信息指导下的特征图局部增强模块，通过统一建模不同粒度的特征，避免了裁剪方法带来的上下文信息损失；2)多尺度特征融合模块，有效聚合多尺度特征图的上下文信息；3)交互关系高级视觉特征生成器，通过学习自动聚焦于增强的细粒度特征，并获取细粒度特征与全局特征之间的长距离依赖，生成丰富的上下文信息和高级视觉特征。在 V-COCO 和 HICO 数据集上验证了本文方法的优越性，其相比基准模型在 V-COCO 数据集上 mAP 提升了 7.7 个百分点、在 HICO 数据集上 mAP 提升了 7.43 个百分点、7.5 个百分点、7.85 个百分点。希望能本文方法够激发人与物体交互检测领域更多的研究灵感。

本文模型在以下两个方面还有提升空间：1)物体的细粒度特征提取还不够精确，未来将探索设计更科学的物体细粒度信息提取方案，增强模型对物体的感知能力，提升模型在复杂场景下识别人-物交互动作的能力；2)提升可变形注意力机制对全局特征的建模能力，可变形注意力机制虽然能很好地建模局部特征，但在对全局依赖关系的建模上还有不足。本团队将继续关注注意力机制在人与物体交互领域的应用。

参 考 文 献

[1] GUPTA S, MALIK J. Visual semantic role labeling[J]. arXiv: 1505.04474, 2015.

[2] SADEGHI M A, FARHADI A. Recognition using visual phrases [C]// 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2011: 1745-1752.

[3] WAN B, ZHOU D, LIU Y, et al. Pose-aware multi-level feature network for human object interaction detection[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 9469-9478.

[4] LI Y L, ZHOU S, HUANG X, et al. Transferable interactiveness knowledge for human-object interaction detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 3585-3594.

[5] YAN Z X, BAI L, LI T S. Lightweight human pose estimation based on self-knowledge distillation and convolution compression[J]. Journal of Chinese Computer Systems, 2024, 45(2): 461-469.

[6] DALAL N, TRIGGS B. Histograms of oriented gradients forhuman detection[C]// Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington: IEEE Computer Society, 2005: 886-893.

[7] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[8] GAO C, ZOU Y, HUANG J B. ican: Instance centric attention network for human-object interaction detect-ion[J]. arXiv: 1808.10437, 2018.

[9] GKIOXARI G, GIRSHICK R, DOLLÁR P, et al. Detecting and recognizing human-object interactions[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 8359-8367.

[10] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]// Advances in Neural Information Processing Systems. 2015.

[11] LI B Z, ZHANG J, WANG B L, et al. Human-Object Interaction Recognition Integrating Multi-level Visual Features [J]. Computer Science, 2022, 49(S2): 643-650.

[12] LIN X, ZOU Q, XU X. Action-guided attention mining and relation reasoning network for human-object interaction detection [C]// Proceedings of the 29th International Conference on International Joint Conferences on Artificial Intelligence. 2021: 1104-1110.

[13] SUN X, HU X, REN T, et al. Human object interaction detection via multi-level conditioned network[C]// Proceedings of the 2020 International Conference on Multimedia Retrieval. 2020: 26-34.

[14] QI S, WANG W, JIA B, et al. Learning human-object interactions by graph parsing neural networks[C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018: 401-417.

[15] WANG H, ZHENG W, YINGBIAO L. Contextual heterogeneous graph network for human-object interaction detection[C]// Computer Vision – ECCV 2020: 16th European Conference. Cham: Springer, 2020: 248-264.

[16] ULUTAN O, IFTEKHAR A S M, MANJUNATH B S. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 13617-13626.

[17] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems. 2017.

[18] ZOU C, WANG B, HU Y, et al. End-to-end human object interaction detection with hoi transformer[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 11825-11834.

[19] TAMURA M, OHASHI H, YOSHINAGA T. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual informa-tion[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 10410-10419.

[20] ZHOU P, CHI M. Relation parsing neural network for human-object interaction detection[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 843-851.

[21] LIU H, MU T J, HUANG X. Detecting human—object interaction with multi-level pairwise feature network[J]. Computational Visual Media, 2021, 7: 229-239.

[22] CHEN M，LIAO Y，LIU S，et al. Reformulating hoi detection as adaptive set prediction[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021：9004-9013.

[23] KIM B，LEE J，KANG J，et al. Hotr：End-to-end human-object interaction detection with transformers[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021：74-83.

[24] PARKJ，PARK J W，LEE J S. Viplo：Vision transformer based pose-conditioned self-loop graph for human-object interaction detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023：17152-17162.

[25] MA S，WANG Y，WANG S，et al. Fgahoi：Fine-grained anchors for human-object interaction detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2024，46(4)：2415-2429.

[26] WU M，GU J，SHEN Y，et al. End-to-end zero-shot hoi detection via vision and language knowledge distillation[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2023：2839-2846.

[27] RADFORD A，KIM J W，HALLACY C，et al. Learning transferable visual models from natural language supervision[C]// International Conference on Machine Learning. PMLR，2021：8748-8763.

DING Yuanbo，born in 2000，postgraduate. His main research interest is recognizing human-object interactionactions.

BAI Lin，born in 1985，associate professor，postgraduate supervisor，is a member of CCF（No. A6951M）. His main research interests include deep learning and computer vision.

（责任编辑：何杨）

关于公布 2025 年度 CCF 推荐教材的通知

CCF 汇聚了计算机领域各个方向的顶尖人才，他们大部分在学校从事本科生和研究生的课程教学或教学管理工作。为鼓励工作在理论/技术前沿的青年教师投身教学工作，创作出丰富的教学成果，由 CCF 教育专委提出动议，成立 CCF 优秀教学成果推荐委员会评选 CCF 推荐教学成果、CCF 推荐教材，设立 CCF 教材出版专项基金。

2025 年 5 月 18 日，首届 CCF 推荐教材开始申报，共收到 52 本教材申请，经过初审、函评、终评和 CCF 优秀教学成果推荐委员会审定，并经过 7 个工作日的公示，现将 2025 年度 CCF 推荐教材公布如下。

教材名	作者	作者单位
软件需求工程方法与实践	金芝，刘璘等	北京大学，清华大学等
数据库系统概论(第 6 版)	王珊，杜小勇，陈红	中国人民大学
数值分析与算法(第 3 版)	喻文健	清华大学
计算机组成与实现	高小鹏，万寒	北京航空航天大学
Python 语言程序设计基础(第 3 版)	嵩天等	北京理工大学
解析深度学习(第 2 版)	魏秀参，杨健	东南大学，南京理工大学
信息安全概论(第 3 版)	李剑	北京邮电大学
人机物融合群智计算	郭斌，刘思聪，於志文	西北工业大学，哈尔滨工程大学
智能计算系统：从深度学习到大模型(第 2 版)	陈云霁等	中国科学院计算技术研究所
离散数学(第 3 版)	屈婉玲，曹永知等	北京大学

据 CCF 微信公众号