目　　录

摘　要

对头颅侧位片的测量分析是现代口腔正畸领域中研究颅面生长发育、畸形诊断与制定矫正方法的关键步骤，分析的结果将直接影响医生对患者基本情况的判断。当对头颅侧位片进行分析时，医生需要在侧位片上标记牙颌、颅面等相关位置的关键坐标点，并定量与定性分析关键坐标点之间的距离以及角度关系，而分析结果也将作为决策与制定诊疗手术方案的重要依据。因此，关键坐标点定位的准确性将直接影响最终的治疗效果。当前临床上大多采用手动定位关键坐标点的方式，导致坐标点定位速度和准确度直接受医生水平影响，标准不统一，经常因为坐标点定位错误导致最终的手术失败。为此，本文将基于深度学习的关键点检测技术应用于口腔正畸的头颅侧位片测量分析中，实现自动化关键坐标点的定位分析，提高定位准确度和时效性。

结合临床需求，并针对手工关键坐标点定位方式存在的定位速度慢、准确度低等问题，本文提出并设计了基于卷积神经网络的头颅侧位片关键点检测定位模型 CenterNetAt（CenterNet-Attention），主要工作及贡献如下：

在模型构建方面：本文将目标检测算法 CenterNet 作为研究的基准网络，该网络通过简化网络结构能够在保持检测精度的同时大幅度减少模型参数量，达到小样本学习的效果。采用改进的深度聚合网络（DLA34）作为基准网络的特征提取模块，主要的改进是将注意力机制引入到网络中，加强网络对头颅侧位片中头骨区域特征的表达能力，减少背景等非特征信息的干扰。之后在该网络上采样前加入可变形卷积，增强模型对个体头骨差异性的建模能力，并将网络中的 Relu 激活函数替换成 LeakyRelu，提高模型的收敛性，保障模型的稳健性。在实验采用的数据集方面：本文将 ISBI 2015 挑战赛的 Cephalometric X-rays 数据集作为原始数据，并运用了直方图均衡化、归一化、锐化等多种数据预处理技术处理该数据集。实验结果表明，本文提出和设计的算法在关键点定位上平均误差达到 1.11mm，优于临床上平均 2mm 以内的可接受误差范围，表明本文所提头颅侧位片关键解剖点自动定位算法能够更加准确地定位关键点，并具有良好的时效性，可以满足临床医学的需求。

在模型应用层面：本文基于所提算法搭建并部署了智能化头颅侧位片测量分析系统，为医生和患者提供了更为全面的视觉信息支持、更加准确的

术前分析以及科学的治疗方案和术后评估,有效缩短了患者临床治疗时间，加强了正畸医疗过程的科学性、智能性。

关键词 口腔正畸；关键点检测；深度学习；CenterNet；注意力机制

ABSTRACT

The analysis of lateral cephalometric radiographs is a cornerstone in contemporary orthodontics for investigating craniofacial growth, diagnosing abnormalities, and planning corrective interventions. The outcomes of these analyses crucially influence a physician's understanding of a patient's condition. When examining lateral cephalometric radiographs, it is imperative for physicians to accurately identify and mark key points related to dental and facial structures, and to assess the distances and angles between these points both quantitatively and qualitatively. These assessments form a vital foundation for decision-making and the formulation of diagnostic and therapeutic strategies, thereby directly impacting the effectiveness of treatment outcomes. Presently, the clinical practice predominantly involves manual identification of these key points, a process that heavily relies on the expertise of the physician, lacks standardization, and is prone to errors, potentially leading to unsuccessful surgical outcomes.

To address these challenges, this paper introduces an application of deep learning-based keypoint detection technology to the field of orthodontics for the measurement and analysis of lateral cephalometric radiographs. This approach aims to automate the localization and analysis of key points, thereby enhancing the accuracy and efficiency of the process.

In light of the clinical demand and the challenges associated with manual localization methods, such as slow speed and potential inaccuracies, this paper proposes a novel convolutional neural network-based model, CenterNetAt (CenterNet-Attention), for the detection and localization of key points on lateral cephalometric radiographs. The model and its contributions are as follows:

From a model construction perspective, this paper leverages the object detection algorithm CenterNet as a foundational network. CenterNet simplifies the network structure to significantly reduce the number of model parameters while maintaining detection accuracy, facilitating learning from small sample sizes. The model incorporates an improved Deep Layer Aggregation network (DLA34) as its feature extraction module, integrating an attention mechanism to enhance the expression of

cranial features in lateral cephalometric radiographs and to minimize interference from non-relevant information such as background noise. Additionally, deformable convolutions are introduced before upsampling to better model the variability in individual cranial structures, and the ReLU activation functions within the network are replaced with LeakyReLU to improve convergence and ensure the model's robustness.

Regarding the dataset used for experimentation, this paper utilizes the Cephalometric X-rays dataset from the ISBI 2015 challenge, employing various data preprocessing techniques like histogram equalization, normalization, and sharpening to prepare the dataset. The experimental findings indicate that the proposed model achieves an average keypoint localization error of 1.11mm, surpassing the clinical threshold of acceptability set at 2mm. This underscores the model's ability to precisely localize critical anatomical points in lateral cephalometric radiographs, thereby meeting the exigencies of clinical practice.

On the application front, the paper details the development and deployment of an intelligent system for the analysis of lateral cephalometric radiographs, built upon the proposed algorithm. This system offers comprehensive visual support, facilitates more accurate preoperative evaluations, and provides scientifically grounded treatment plans and postoperative assessments for both physicians and patients. By significantly reducing the duration of clinical treatments and enhancing the scientific and intelligent quotient of orthodontic procedures, this system represents a significant advancement in the field.

Key words: orthodontics; keypoint detection; deep learning; CenterNet; attention mechanism

第一章 绪论

1.1 研究背景与意义

世界卫生组织（WHO）认为口腔健康是整体健康的重要组成部分，并将其列为人类健康的十大基本标准之一。随着 2017 年《中国防治慢性病中长期规划（2017-2025 年）》的颁布，口腔疾病被正式纳入慢性病的管理范畴。到了 2019 年，国家卫生健康委员会再次强调了口腔健康的重要性，并发布了《健康口腔行动方案（2019-2025 年）》，标志着口腔健康受到的社会关注已经达到前所未有的高度。这一系列国家级政策的实施，不仅反映了公众对口腔疾病的持续关注，同时也侧面说明人们对个人形象及美容意识的增强。因此，随着对口腔健康及其美学[1-2]需求的持续增长，口腔正畸[3-4]领域的研究和实践也面临着越来越高的期待。作为一个专业的医学分支，正畸学[5]主要研究和处理口腔颌面发育的异常情况及其矫正治疗。它关注的是牙齿、颌骨、软组织的结构及其相互作用，旨在通过对这些结构的细致研究和精准诊断，采取有效的治疗方案，如使用矫正器或手术等方法，以提高患者的咬合效能、面部外观和口腔健康。所以，口腔正畸不仅关乎功能的改善，更涉及面部美学及个体心理健康的提升。而头颅侧位片关键坐标点定位（如图 1.1 所示）是正畸的关键步骤。在口腔正畸治疗的前期，医生们主要通过定位并分析头颅侧位片中的关键坐标点，并根据分析结果评估患者颌面结构的生长发育状况，以此作为后续采取对应治疗措施的主要依据。所以，精准的头颅侧位片关键解剖点的定位对于制定合理的治疗计划至关重要，它直接关系到治疗效果的优劣和患者最终的满意度。在当前临床实践中对头颅侧位片中关键坐标点的定位分析通常由手工完成，导致关键点定位存在精度差，耗时长，易受主观性影响等问题。

随着计算机硬件性能的提升和大规模数据集的可用性，使得深度学习得到了蓬勃的发展，并在图像处理和模式识别等领域取得了显著的成就[6-10]，这为口腔正畸中头颅侧位片中关键坐标点的定位研究提供了新的方向。因此，将深度学习相关技术应用于头影关键点的定位分析领域，有助于解决当前头影关键坐标点定位在临床实践中存在的问题，同时帮助医生更全面、准确、快速的分析患者颅颌面部结构、给患者提供更为科学有效的治疗方案、缩短医疗过程，提高医生工作效率。

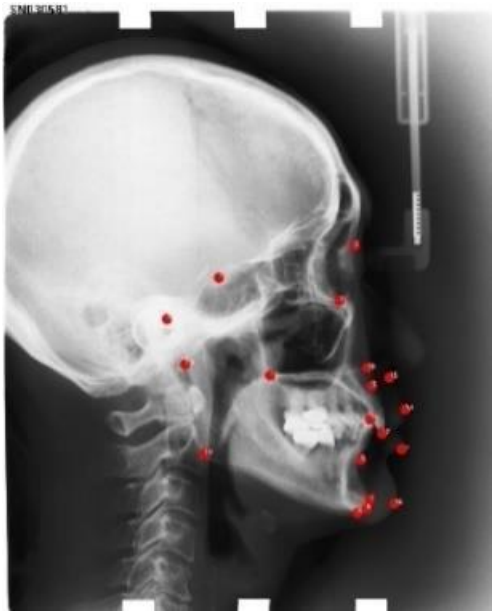


图 1.1 头颅侧位片关键点标记图

1.2 医学解剖点

头颅侧位片中的解剖关键点在构建测量线、平面以及角度方面起着至关重要的作用（如表 1.1 所示）[11]，医生通过这些精确的测量来评估患者的面部生长发展状况，诊断可能存在的异常，并据此制定出相应的治疗计划。在正畸学、颌面外科学以及其他相关的医疗领域，标准的头影测量分析通常包括从几十个到数十个不同的解剖关键点，覆盖颅骨、面部软组织以及牙齿等关键区域。对于那些要求更为复杂或更加细致的分析，可能会涉及更多的解剖点，并扩展到更广泛的头部区域。在日常的临床工作中，这些关键点大致可以分为两大类：第一类是牙齿和颅骨结构定位用的解剖点；第二类则包括通过头影图上的特征点间的计算得出的，如测量平面交点等派生点。这些关键点的精确分析对于确保治疗方案的科学性和有效性至关重要。本文以 19 个解剖关键点作为研究的基准点（其在头影中的位置如图 1.2 所示），具体如下：（1） 骨性解剖点 13 个（Skeletal Landmarks）: 蝶鞍点（1/sella/S）、鼻根点（2/Nasion/N）、耳点（4/Porion/Po）、眶点（3/orbitale/Or）、上齿槽座点（5/Subspinale/A）、颏下点（8/Menton/Me）、下齿槽座点（6/Supramental/B）、颏顶点（9/Gnathion/Gn）、颏前点（7/Pogonion/Pog）、下颌角点（10/Gonion/Go）、关节点（19/articulare/Ar）、后鼻棘点（17/posterior nasal spine/PNS）、前鼻棘点（18/anterior nasal spine/ANS）。（2）牙齿解剖点 2 个（Dentak Landmarks）：上中切牙点（12/upper incisor/Ui）、下切牙点（11/lower incisor/Li）。（3）软组织解剖点 4 个：

上唇突点（13/Labrale Superius/Ls）、下唇突点（14/Labrale Inferius/Li）、鼻下点（15/subnasale/Sn）、软组织之颏前点（16/pogonion of soft tissue/Pos）。

表 1.1 部分测量项目及意义

类目	测量项目	Mean	SD	意义
骨骼	SNA	81.69	2.54	上颌骨基骨相对于颅底的前后位置关系
	SNB	78.94	2.19	下颌骨基骨相对于颅底的前后位置关系
	ANB	2.75	1.16	上下颌基骨间的前后向位置关系
	S-Ptm	17.98	2.98	上颌骨后缘相对于前颅底的前后向位置关系
	A-Ptm	44.89	2.76	上颌骨基骨的长度
面高	N-ANS	57.73	3.88	面中份高度
	ANS-Me	70.47	4.3	面下份高度
	S-Go	83.94	4.7	后面高
牙齿及牙槽骨	U1 to L1	123.2	6.18	上下中切牙的倾斜度
	U1 to SN	74.94	6.22	上中切牙的唇轴度
软组织侧貌	UL-EP	-0.46	1.92	上唇突度
	LL-EP	1.31	1.92	下唇突度
	Z-Angle	74.06	4.57	Z 角 颏部突度

表中的 Mean 表示平均值，SD 表示允许的误差

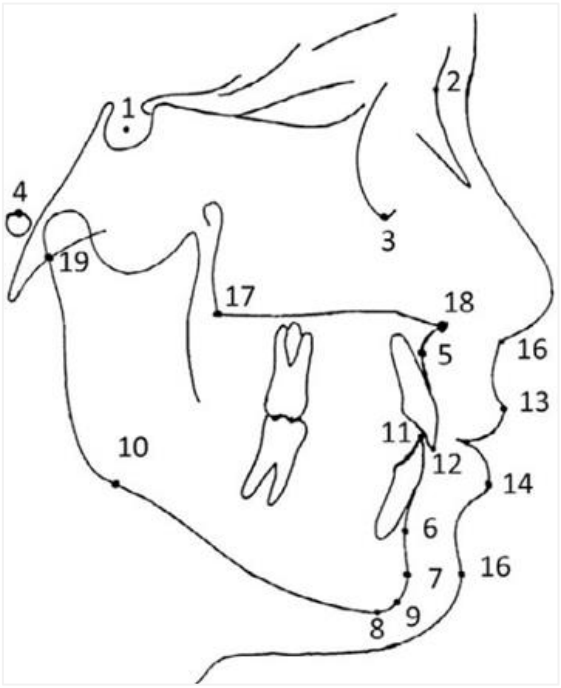


图 1.2 19 个关键点示意图

1.3 头颅侧位片关键解剖点检测算法研究现状

头颅侧位片关键解剖点检测算法（多数研究中也称头影测量算法）是口腔正畸和颅面分析领域的重要研究方向，旨在通过自动化技术准确识别和定位头颅侧位片中的关键解剖点，从而辅助医生进行诊断和治疗规划。自 20 世纪 20 年代研究人员提出头颅侧位片测量分析以来，相关技术得到了快速发展，并广泛应用于口腔正畸及颌骨矫正中。

首先基于手动头颅侧位片测量分析的方式主要是医生使用直尺、量角器等测量工具在玻璃纸或醋酸纸上人工画出侧位片中的颅颌面轮廓线等（如图 1.3 所示），然后根据使用的分析方法确定所需关键点，并定量和定性分析相关测量项。然而手动测量方法需要医生亲自在侧位片上用直尺和量角器等工具进行标记和测量，这一过程十分耗时，尤其是在高负荷的医疗环境中，这种方法的效率难以满足临床需求，且手动测量的精确度高度依赖于操作者的经验和技巧，不同医生之间可能存在明显的测量差异，甚至同一医生在不同时间的测量结果也可能有所不同，这种主观性和不一致性可能影响诊断和治疗计划的准确性。同时由于测量过程涉及多个步骤，每一步的小差错都可能在最终结果中累积放大，导致测量的重复性较差，这对于需要多次测量以跟踪治疗进展的情况尤为不利。

自 20 世纪 80 至 90 代，随着数字图像技术的发展和进步，研究人员开始将相关技术应用于头影侧位片的测量分析，旨在提高头影解剖关键点的定位准确性，并降低观测者的主观判断的影响。例如 Grau 等人[12]采用基于先验知识的边缘检测法[13-15]，而 Keustermans 等人[16]采用基于局部外观模型的模板匹配算法。同时，Yue 等人[17]采用基于形状的主动形状模型等多种处理技术，以准确定位头影解剖关键点的定点检测。尽管这些方法在某种程度上提高了定位精度，但它们各自也存在明显的局限。例如，基于先验知识的算法仅在使用高分辨率照片时才会有比较好的效果，且由于使用了预定义线，角等先验知识，这些模型的泛化能力受限，只对特定结构有效。基于模板匹配的算法对解剖点附近的灰度变换较为敏感，在实验过程中会出现偏离结果的现象。而基于形状的模型在处理遮挡部分或形状变化时存在困难。为了克服这些局限，Kafieh 等人[18-20]在 Yue 等人的研究基础上，将神经网络与主动形状模型相结合解剖定位头影特征点。此外，Rudolph 等人[22]提出空间光谱学定位算法，通过数学处理将头影图像分成 75 个不同的特征值，然后根据每个像素的特定值计算属于某个关键点的概率 [21]。这些创新方法虽然在提升解剖点定位性能方面取得了一定的成果，但面临着对噪声和图像变形较为敏感、以及在处理复杂背景时算法准确度不足等挑战。

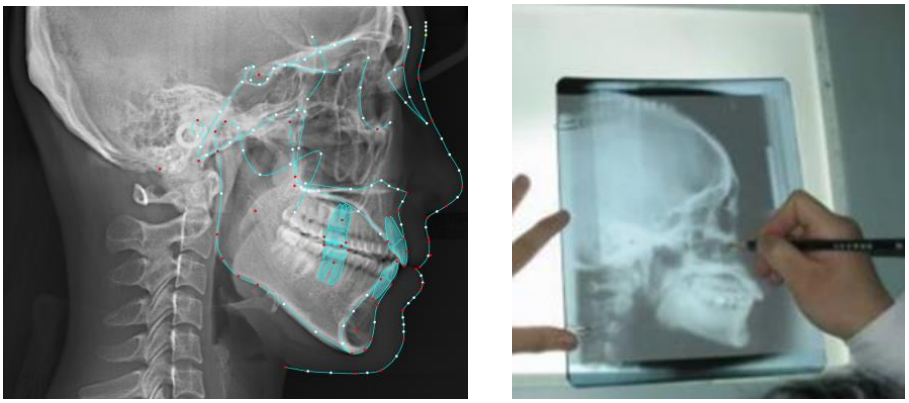
近年来，随着机器学习在图像识别、处理及分析等领域的广泛应用[22-23]，并具有优秀的性能。研究人员开始探索将机器学习算法应用于头影图像解剖关键点的自动检测中，以期进一步提升算法性能。例如，Ibragimov 等人[24]提出将 Haar-like 特征提取与随机森林分类器相结合的检测算法，该算法利用 Haar-like 特征在捕捉图像中的边缘、线条和其他简单纹理信息方面的高效性能与随机森林分类器相结合，有效提高了解剖特征点的定位精度。Vandaele 等人[25]采用了一种集成学习策略，该策略通过构建多个决策树，并利用基于像素的多层次特征方法对于图像的局部和全局信息的关注训练检测模型。Chu 等人[26-27]将稀疏形状结构应用于随机森林分类器构建检测模型。Mirzaalian 等人[28-29]通过在随机森林分类器中加入通用特征提取算法定位解剖点。Linder 等人[30-31]提出将基于随机森林的回归投票方法和局部条件模型相结合自动定位解剖特征点。尽管这些方法与传统图像处理算法相比，在检测精度以及效率方面取得了很大的进步，但基于传统机器学习算法的头影解剖点定位往往依赖于人工设计的特征，无法做到端到端的定位检测，且在处理高维数据（如图像和视频）时，传统机器学习算法往往出现“维数灾难”的现象，模型性能可能会下降。

伴随着深度学习在图像分割分类、物体检测、目标跟踪等领域的成功应用[32-34]，研究者开始将深度学习相关技术应用到口腔正畸领域，实现端到端的关键点检测，以期进一步提高检测的准确度和效率。例如，Arik 等人[35]以 CNN 为基础，并融合统计形状模型成功构建了头影关键点定位算法。Lee 等人[36]通过将所有标注点的 x 和 y 坐标转化为 38 位的一维向量，构建了基于 CNN 的多元回归系统来预测坐标变量。Zhusi Zhong 等人[37]提出了一个深度编码器-解码器模型，该模型基于两级 U 型网络[38]，将全局标志点位置与局部高分辨率特征响应相结合，对多通道热图进行回归，实现标志点检测。Jiahong Qian 等人[39]根据头影图像的特点以及标志点的分布特点，提出了基于 R-CNN[40]的头影标志点快速检测模型 CephaNet。Dai 等人[41]提出了一种使用生成性对抗网络（GAN）[42]的头影测量标记方法，利用 GAN 模型为每个标志点生成距离图，通过对相应距离地图进行回归投票得到对应关键点的坐标。Chen 等人[43]提出了一种与注意力机制相结合的特征金字塔融合模块，通过对不同层次的特征进行融合，并将偏移图与热图进行组合，最终通过逐像素投票来获得标志点的位置。Yue[44]等人通过将卷积编码器和带跳跃连接的解码器组合起来构建头影解剖特征点定位。任家豪等人[45]通过采用预训练的 MobileNetV2 提取特征, 在解码网络中引入由粗到细的中间监督，并将得到的热图与特征图融合定位头影关键解剖点。

最近几年锥形束计算机断层摄影技术（Cone Beam Computed Tomography，CBCT）快速发展起来，基于 CBCT 医学图像的三维头影测量图像解剖特征点的自动定位方法逐渐受到研究者的关注，并且因其高效的重建速度和较低的电离辐射水平，在口腔医学的诊断和治疗等领域获得了广泛应用[46-48]。CBCT 技术的主要优势在于其能够提供高分辨率的三维图像，这些图像不仅能够捕捉颅颌面结构的详细信息，而且允许进行精确的三维测量。并且与二维 X 线图像相比，CBCT 图像通过消除重叠和扭曲，避免了二维图像中常见的固有缺陷，使得测量结果更接近实际情况。此外，CBCT 技术还使得医生可以在三维空间内测量任意两点之间的距离，极大地丰富了颅颌面结构分析的内容，并提高了测量的准确性。

现有方法不足和挑战：借助机器学习和深度学习技术，头影侧位片关键解剖点自动定位技术取得了显著进展，其定位准确性不断提升。然而，由于个体之间头颅和颌骨结构的显著差异，加之成像设备的多样性带来的图像质量不一致，获取大规模、高质量、精确标注的头影侧位片图像数据集变得极为困难。这限制了深度学习模型的训练和泛化能力，影响了定位技术的普遍适用性和准确性。同时，为了提高

定位的准确性，现有的深度学习模型往往结构复杂，参数众多，导致模型尺寸较大。这不仅增加了计算资源的需求，也提高了模型的部署成本。而现阶段基于 CBCT 的三维图像处理方式，同样存在与上述类似的不足。因此，发展新型的基于深度学习的头影关键解剖点自动定位方法具有重要的研究意义和临床应用价值。



（a）头颅侧位片轮廓线图　　（b）手动标记解剖点

图 1.3 手动标记头颅侧位片关键点示意图

1.4 本文主要研究内容和组织结构

头颅侧位片中关键解剖点的定位分析，对正畸治疗具有重要的影响。定位结果直接决定临床医生治疗方案的选择，以及后续患者的治疗效果。随着深度学习的不断发展，并在医学图像处理等领域的广泛使用，使得构建速度快，准确度高的头影关键点检测算法成为可能。然而当前基于卷积神经网络的头影关键解剖点定位研究存在样本不足、成像设备不同带来的图像质量参差不齐、个体之间头骨形状复杂多样、小样本学习下模型泛化能力差等难题。基于此，本文利用深度学习相关知识，提出了基于注意力机制的头颅侧位片关键解剖点自动定位算法，主要内容如下：本文以目标检测算法 CenterNet[49]作为研究的主要框架；深层聚合网路[50]（Deep Layer Aggregation，DLA）作为特征提取网络，并在其中引入注意力机制，利用注意力机制对关键特征的关注度，增强网络的特征表达能力，减少背景信息干扰。同时，将可变形卷积[51]（Deformable Convolutional Networks，DCN）加入到特征提取网络中，利用可变形卷积对于形态变化的学习能力，提升模型对于个体头骨差异性和局部特征信息的建模能力。在实验中，本文使用 ISBI 2015 挑战赛上的 Cephalometry X-rays 数据作为训练测试验证模型的主要数据源，并运用了多种数据预处理技术处理数据集。实验表明，本文所提算法在头颅侧位片关键标志点检测任务中提高了检测精度，模型具有优越的性能。之后利用本文研究成果搭建并部署了基于头颅侧位片关键解

剖点定位技术的智能口腔正畸平台，有力提高了医生工作过程的科学性以及效率。以下是本文主要的章节安排：

第一章：绪论。首先介绍了本课题研究的背景与意义，并详细介绍了解剖点；接着介绍了头颅侧位片关键解剖点定位检测技术的发展现状。最后列举了本文主要研究内容及章节安排。

第二章：相关理论与技术研究。本章主要阐述了本文用到的神经网络理论知识以及相关技术，包括感知器、多层神经网络中的激活函数、梯度下降算法以及卷积神经网络中的卷积、池化、全连接以及损失函数等，并介绍了几种不类型的卷积；最后对章节内容进行了总结。

第三章：设计基于注意力机制的头颅侧位片关键解剖点的定位检测算法。首先对本文使用的 CenterNet 和 DLA34 网络进行了详细的介绍，并对其中的改进做了说明；接着介绍了本文的数据集以及基于数据集做的一系列预处理操作；然后对本文实验所需的环境配置及参数设计进行了介绍；之后对实验结果进行了分析；最后对章节内容进行了总结。

第四章：搭建基于头颅侧位片关键解剖特征点定位检测技术的口腔正畸平台。本章节主要介绍了用于智能化口腔正畸的原型系统设计，包括系统结构、数据库设计、开发环境的准备、系统主要功能介绍以及系统评价等内容。

第五章：总结全文。对本文所提出的算法进行回顾，分析算法存在的问题，并对本文所提算法的进一步改进作出展望。

第二章　相关理论与技术研究

2.1 神经网络理论

神经网络是一种受生物神经网络（如人脑）启发的数学模型，用于执行机器学习任务。它由大量的单元（称为“神经元”）组成，这些单元按层次排列并通过加权连接相互作用。神经网络利用一系列输入数据进行训练，通过调整神经元之间连接的权重来学习数据的内在模式和关系，以此来实现分类、回归、聚类、模式识别等多种复杂任务。其主要思想是网络中的每个神经元接收来自前一层神经元的加权输入，将这些输入值汇总，并通过一个非线性激活函数处理后产生输出，这个输出随后可以作为下一层神经元的输入。整个网络的第一层称为输入层，接收外部数据；最后一层称为输出层，提供任务的结果；位于输入层和输出层之间的是隐藏层，它们负责处理复杂的数据特征提取和转换。神经网络的训练过程涉及前向传播和反向传播两个阶段。在前向传播阶段，输入数据在网络中向前传递，直到生成输出。然后，根据损失函数（即网络输出与实际目标之间的差异度量）计算误差。在反向传播阶段，误差被反向传递回网络，以便通过梯度下降或其他优化算法调整权重，从而最小化损失函数，最终找到最优参数下的模型。随着深度学习的发展，出现了多种神经网络结构，包括用于图像处理的卷积神经网络（CNN）、用于处理序列数据的循环神经网络（RNN）以及用于捕获长距离依赖关系的长短期记忆网络（LSTM）等。图 2.1 展示了神经网络的发展历程。

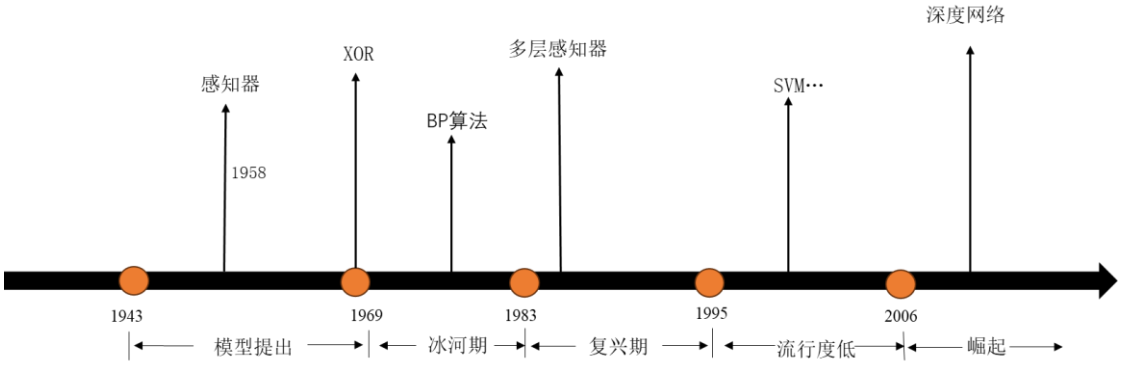


图 2.1 人工神经网络发展脉络

2.2 感知器原理

感知器是最早期的人工神经网络结构之一，由美国心理学家兼计算机科学家 Frank Rosenblatt 于 1957 年创立。作为一种基础的神经网络模型，感知器只包含单一的人工神经元，能够处理多个输入信号。每个信号在输入前会被相应的权重加权，所有加权输入随后相加得到一个总和。之后总和会通过一个特定的激活函数进行转换，产生一个二元输出。该输出通常用于区分两个不同的类别，从而使感知器成为有效的线性二分类工具，其工作流程如图 2.2 所示，数学形式如公式 2.1 所示。

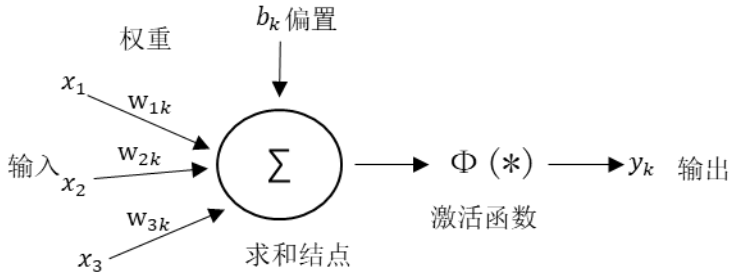


图 2.2 单层感知器结构图

$$y_k = \Phi(\sum_{i=0}^{n} w_{ik} x_i + b_k) = \Phi(w^T x + b) \tag{2.1}$$

其中x_i表示输入信号；y_k为输出信号；w_{ik} 表示x_i相对于y_k的权重；b_k表示神经元的偏置值；$\Phi(*)$表示激活函数，通常是 Softmax 等。Softmax 公式如 2.2 所示：

$$\sigma(z)_j = \frac{e^{z_j}}{\sum k^{e^{zk}}} \tag{2.2}$$

Softmax 函数将一个K 维的任意实数向量映射成另一个K维的实数向量，其中新的向量中的每个元素取值都介于（0，1）之间，并且所有维度模长之和为 1，因此经常用作多分类任务的最后一层，表示输入是某个类别的概率。例如经过 Softmax 函数后某一输出值为 0.9，则在分类任务中表示属于某个类别的概率为 90%

2.3 多层神经网络

单个感知器由于其固有的结构限制，仅能够处理具有线性决策边界的二元分类任务，难以应对复杂的非线性数据分布。受人脑处理信息方式的启发，当将多个感知器按照一定结构相互连接组成网络时，这种组合便能够显著扩展模型的表达能力，使其能够捕捉并表示丰富多样的非线性关系。这种由众多感知器构成的复杂网络，

被广泛称为多层神经网络，基本结构如图 2.3 所示。在图 2.3 中，多层神经网络主要包括一个由 5 个输入信号组成的输入层，2 个完全连接的隐藏层，其中第一个隐藏层与输入层连接，以及一个与第二个隐藏层完全连接的输出层构成。多层神经网络在单个感知器的基础上有以下几个方面的改进：1.加入隐藏层，并且允许有多个隐藏层。2. 增加输出层的神经元的个数，可以有多个输出，使得模型能灵活的应用于分类回归，以及其他的学习领域。3. 对激活函数做扩展，包括使用 Sigmoid 函数、Relu 等。

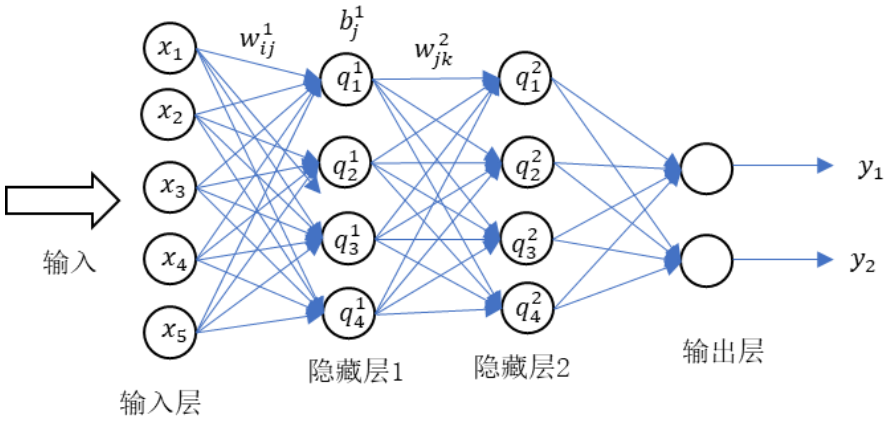


图 2.3 多层神经网络

如上图所示，假设有一组数据 X，网络具有两个隐藏层，并通过全连接的方式前后相连。w_{ij}^1表示输入层第 i 个神经元与第一个隐藏层的第 j 个神经元的连接权重，偏置为b_j^1表示第 1 个隐藏层第 j 个神经元的偏置项，以此类推，那么网络的最终输出 Y 为：

$$Q^1 = f^1(w^1 \cdot X + b^1) \tag{2.3}$$

$$Q^2 = f^2(w^2 \cdot Q^1 + b^2) \tag{2.4}$$

$$Y = w^y \cdot Q^2 + b^y \tag{2.5}$$

其中Q^1表示隐藏层 1 的输出，$f^1(.)$表示应用于隐藏层 1 的激活函数，Q^2表示隐藏层 2 的输出，$f^2(.)$表示应用于隐藏层 2 的激活函数。在实际使用中，通常在最后一层使用 Softmax 等激活函数将输出结果离散化，表示最终分类结果的概率分布。

2.3.1 激活函数

激活函数（Activation Functions）对于神经网络模型去学习、理解非常复杂和非线性的函数来说具有十分重要的作用。如果不使用激活函数，则输出信号将仅仅是一个简单的线性函数，而线性函数的复杂性有限，能从数据中学习复杂函数映射的能力更小，因此将激活函数引入到神经网络中有利于处理非线性等复杂问题，且函数可导的激活函数可以直接利用数值优化的方法来学习网络参数。在设计激活函数时，函数及其导函数要尽可能的简单，有利于提高网络计算效率，并且激活函数的导函数的值域要在一个合适的区间内，不能太大也不能太小，否则会影响训练的效率和稳定性。下面介绍几种常用的激活函数，主要有 Sigmoid、Tanh、Relu、LeakyRelu 等。

Sigmoid 函数将任意实数映射到（0, 1）之间，因此常被用作输出层的激活函数，特别适用于二分类问题中，将模型输出解释为样本属于某个类别的概率。Sigmoid 函数具有平滑性，对于输入的小变化有连续的输出变化，这使得它在反向传播算法中的梯度计算相对简单。但当 Sigmoid 函数的输入值非常大或非常小时，函数的梯度接近于零，这可能导致梯度消失的问题，使得训练过程变得缓慢。Sigmoid 函数图像如图 2.4 所示，数学表达式为：

$$\sigma(x)=\frac{1}{1+e^{x}} \tag{2.6}$$

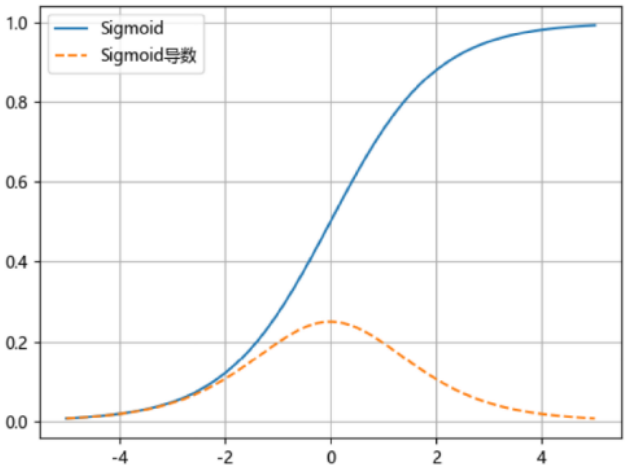


图 2.4　Sigmoid 函数图

Tanh 函数范围为-1 到 1 之间，相较于 Sigmoid 函数有更为广的范围，并具有平滑性和连续性，适用于梯度下降等优化算法。Tanh 函数在原点附近接近线性，使得其相对于 Sigmoid 函数来说，在梯度下降中更容易学习。Tanh 的输出均值接近零，

有助于中心化数据，减小网络的偏移。虽然 Tanh 函数相对于 Sigmoid 函数缓解了梯度消失问题，但在深度神经网络中仍然可能存在该问题，尤其是网络层数较多时。Tanh 函数图像如图 2.5 所示，数学表达式为：

$$\mathrm{Tanh}(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{2.7}$$

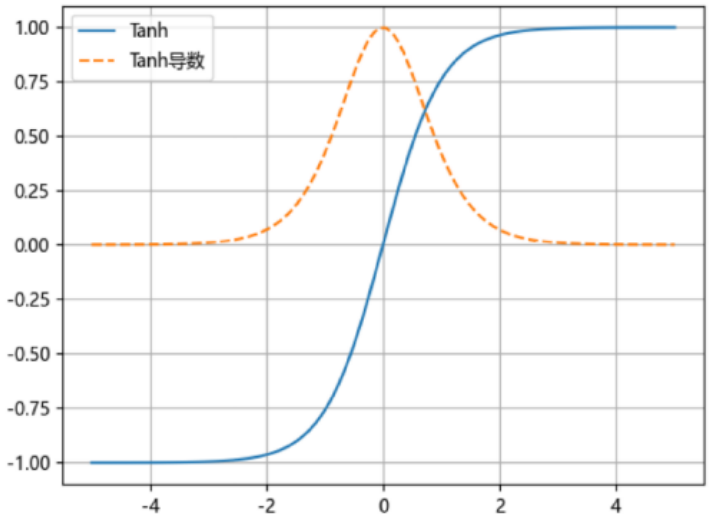


图 2.5 Tanh 函数图

Relu 是一种非线性激活函数，能够引入非线性变换，使神经网络具备学习非线性关系的能力。Relu 函数当输入小于 0 时，输出为 0；当输入大于等于 0 时，输出等于输入。这种特性使得网络中的很多神经元在训练过程中变得稀疏激活，即只有部分神经元被激活，有助于提升模型的泛化能力，并且由于不涉及复杂的数学运算，使得 Relu 激活函数的计算非常高效，且反向传播的计算相对简单。但 Relu 函数在负数区域的梯度为 0，可能导致梯度爆炸问题，即在反向传播过程中一些神经元的权重更新过大。而当输入为负数时，Relu 输出为 0，这可能导致神经元失活，称为“Dead Relu”问题，即某些神经元在训练过程中无法激活。Relu 函数图像如图 2.6 所示，数学表达式为：

$$\mathrm{Relu}(x)=\begin{cases}x, & x\geq 0\\ 0, & x<0\end{cases} \tag{2.8}$$

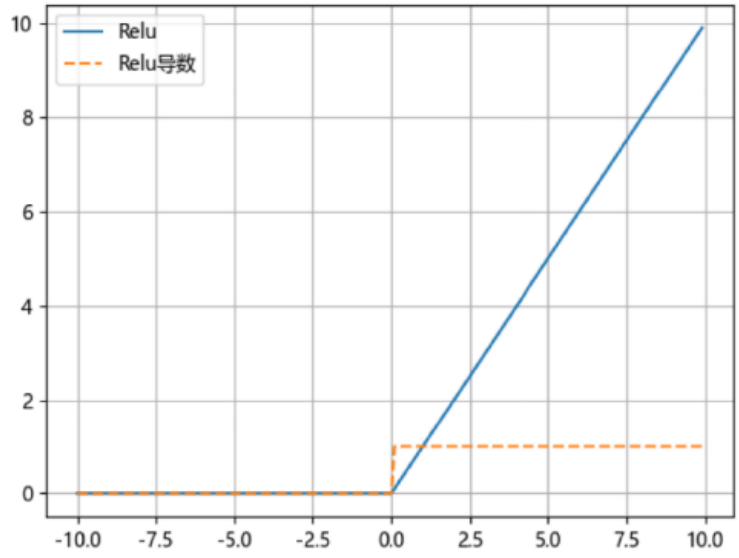


图 2.6 Relu 函数图

LeakyRelu 函数是对 Relu 函数的改进，它的输出在（−∞,+∞）范围内。LeakyRelu 在 Relu 的基础上引入了一个小的负斜率，使得负数输入时也有非零输出，从而解决了传统 Relu 可能导致的“Dead Relu”问题。LeakyRelu 函数图像如图 2.7 所示，数学表达式为：

$$\text{LeakyRelu}(x)=\begin{cases}x, & \text{if } x\geq 0\\ yx, & \text{if } x<0\end{cases} \tag{2.9}$$

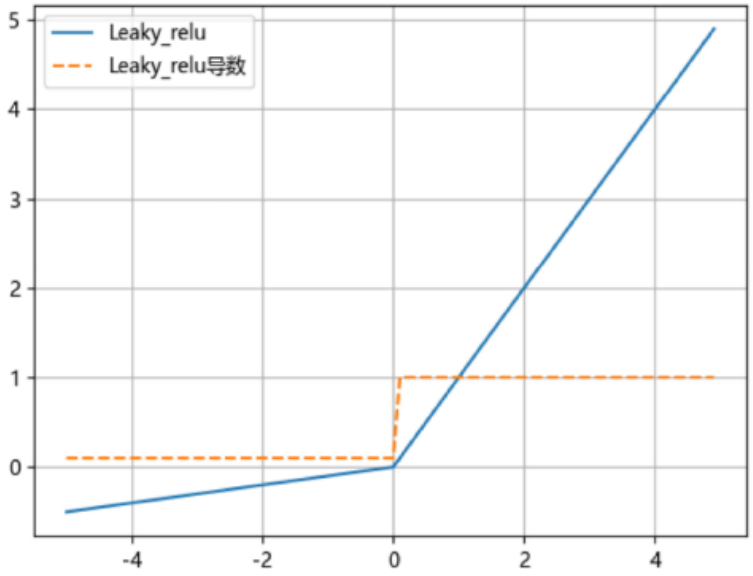


图 2.7 LeakyRelu 函数图

2.3.2 梯度下降算法

在神经网络的训练过程中，想要寻找到最小损失函数下的网络参数，主要是通过梯度下降法来求解。梯度下降法通过不断沿着目标函数的梯度方向进行迭代，以寻找使目标函数达到最小值的参数值，其过程可用 2.10 数学公式表示。

$$\theta_{new} = \theta_{old} - \eta \nabla_{\theta} J(\theta) \tag{2.10}$$

其中θ_{old}是当前迭代的参数值；θ_{new}更新后的参数值；η 为学习率(Learning Rate, LR)，是一个超参数，主要用于控制每次更新中参数移动向最小损失函数值的步长大小；θ 表示模型参数向量；$J(\theta)$表示损失函数；$\nabla_{\theta} J(\theta)$是损失函数关于参数 θ 的梯度，表示损失函数在 θ 处的斜率或者上升方向。该公式的具体意思是，为了使目标损失函数减小，沿着梯度的反方向以步长η更新参数。通过不断迭代这个更新过程，可以逐步优化模型参数，使得目标函数达到最小值。当前梯度下降法主要有以下 3 种不同的策略：

(1) 批量梯度下降（BGD）：在每一次迭代中，使用全部训练样本来计算梯度。批量梯度下降法收敛速度相对较稳定，能够保证在凸函数情况下达到全局最优解。但当训练样本较大时，计算每次迭代的梯度可能非常耗时。

(2) 随机梯度下降（SGD）：每次更新的时候只考虑了一个样本点。随机梯度下降法能够大大加快训练速度，但由于每次迭代仅使用单个样本，使得更新方向的估计可能存在较大的方差，导致收敛过程不够稳定，可能会偏离最优解，且 SGD 对噪声也更加敏感。

(3) 小批量梯度下降（MBGD）：在每一次迭代中，使用一小批量（通常称为 batch-size）的样本来计算梯度。小批量梯度下降法通常是介于批量梯度下降和随机梯度下降之间的折中方案，兼具了批量梯度下降和随机梯度下降的优点，计算速度较快且稳定性较好。但需要调节批量大小，过大会增加计算开销，过小可能会引入噪声。

综上可知，不同的梯度下降适合不同的任务。在综合考虑任务、数据集特点、实验软硬件设备条件等因素后，本文选用 MBGD 梯度下降法优化模型。

2.4 卷积神经网络

在 20 世纪 60 年代，神经生理学家 Hubel 和 Wiesel 的研究揭示了猫的视觉皮层神经元如何响应视觉刺激的局部敏感区域和特定方向。他们发现这些神经元以一种层次化的方式组织起来，较低层次的神经元响应简单的刺激，而较高层次的神经元

则能够对更复杂的模式做出反应。这一发现不仅为理解视觉处理的生物学基础奠定了重要基石，也为人工神经网络的设计提供了灵感，特别是在设计能够高效处理视觉信息的网络结构方面，这些原理成为了后来卷积神经网络（Convolutional Neural Networks，CNN）的基础。卷积神经网络是一种专门用于处理具有已知网格结构数据的深度学习模型，如时间序列数据和图像数据，主要由输入层、卷积层、池化层、全连接层以及输出等组成，并且卷积层、池化层、全连接层的数量不固定。如图2.8展示了卷积神经网络基本架构，其基本工作思路是通过周期性的使用卷积操作反复的自动提取数据的特征，这些卷积操作具有可学习的过滤器，能够捕捉到输入数据中的局部模式，然后通过全连接层整合提取到的特征信息，最后经过 Softmax 等激活函数输出模型预测结果。由于该网络可以直接将原始图像作为输入，避免了对图像的复杂前期预处理，因而在图像处理、计算机视觉等领域得到了广泛的应用。

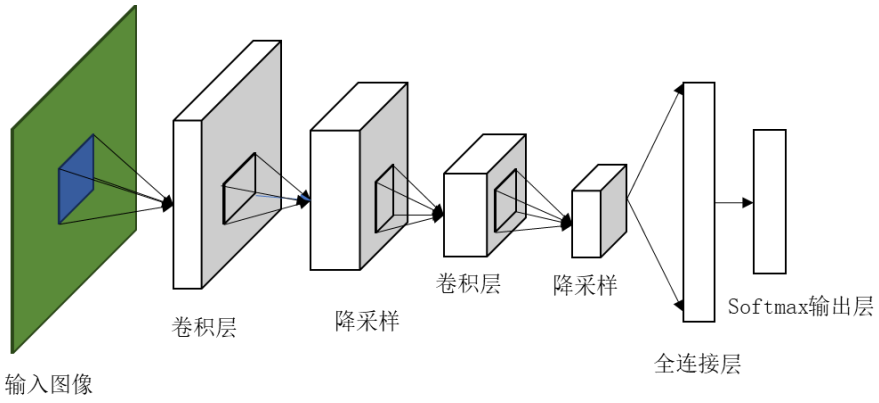


图 2.8 卷积神经网络结构图

2.4.1 卷积层

作为CNN网络的核心，卷积层主要负责从输入图像中提取特征。在卷积层中，小的卷积核（或过滤器）在输入图像上滑动，通过计算卷积核和图像的局部区域之间的点积来生成特征图（Feature Maps）。通过卷积能够学习到图像中重要的视觉特征，如边缘、角点、纹理等，并且卷积操作能够降低参数量。下面主要展示了卷积运算涉及的概念、基本运算过程以及不同类型的卷积。以图像类型的卷积为例，卷积核的数据形状为$[H_c, W_c, C_c]$，其中H_c指卷积核高度，W_c指卷积核宽度，C_c为卷积核的特征通道数量且$C_c \leq C$。在卷积层中被训练的参数主要是卷积核的参数和偏置参数。其运算过程如图 2.9 所示。假设输入数据尺寸 5×5，卷积核 1 个，卷积核尺寸 3×3，特征图尺寸 3×3。

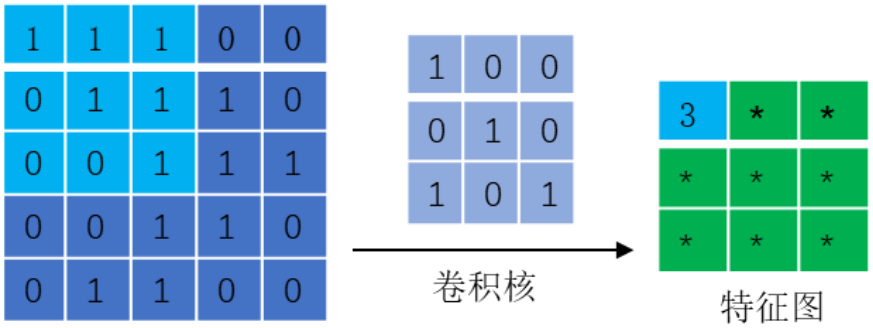


图 2.9 卷积运算

2.4.2 池化层

池化（Pooling）是指对每个区域进行下采样（Down Sampling）得到一个值，作为这个区域的概括。池化的作用是降低特征图参数量、保留图像显著特征、降低过拟合、提高模型的泛化能力。有时输入图像太大，要减少训练参数的数量，可以在卷积层之间周期性地引进池化层。池化层一般分为最大池化（Max Pooling）和平均池化（Mean Pooling）。如图 2.10 和 2.11 所示，展示了 4×4 的特征图经过最大池化和平均池化后的结果。

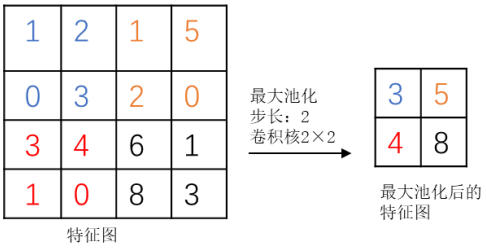


图 2.10 最大池化

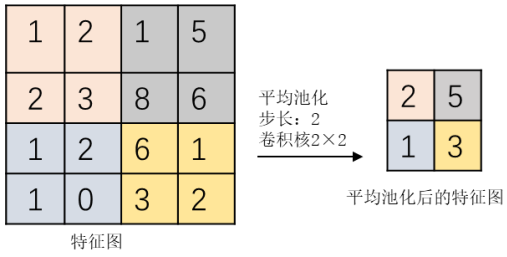


图 2.11 平均池化

2.4.3 全连接层

全连接层（Fully Connected Layers，简称 FC）用于整合提取的特征，将通过卷积和池化得到的特征映射到线性可分的空间。通常全连接层处理方式为将最后一个卷积层的特征图展开平摊成一个长的列向量，然后将所有特征传递给具体的分类器进行分类或者回归处理，一般使用 Softmax 激活函数将最终的输出量化。在整个卷积神经网络中全连接层起到“分类器”的作用。其结构如图 2.12 所示。

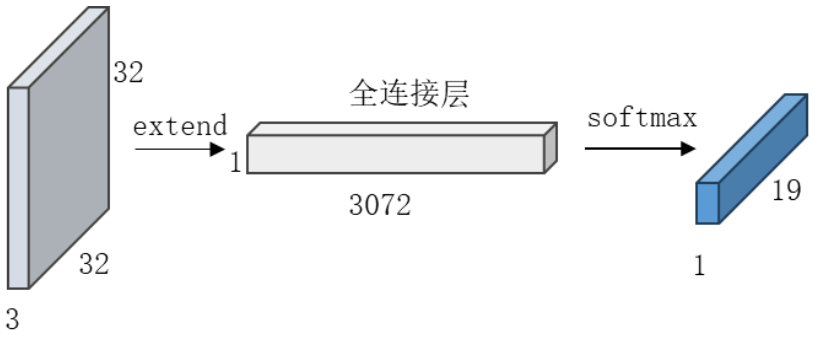


图 2.12 全连接示意图

2.4.4 损失函数

损失函数（Loss Function）是全连接层实现分类的重要一步。损失函数衡量了网络输出的预测值与真实样本标签的距离。神经网络训练的主要目标是最小化在训练集上的损失函数。

本文主要使用加权损失和作为训练的总损失函数评估模型的效果，其定义如下：

$$\mathrm{L}_{\mathrm{tot}} = L_{hk} + \lambda_o L_o \tag{2.11}$$

其中L_{hk}为热图损失（Heatmap Loss），L_o目标偏移量损失（Reg Loss）。其中热图损失公式为：

$$\mathrm{L}_{\mathrm{hk}} = -\frac{1}{n}\begin{cases}(1-\hat{Y}_{xyc})^{\alpha}\log(\hat{Y}_{xyc}) & \text{if } \mathrm{Y}_{xyc}=1\\ (1-Y_{xyc})^{\beta}(\hat{Y}_{xyc})^{\alpha}\log(1-\hat{Y}_{xyc}) & \text{otherwise}\end{cases} \tag{2.12}$$

N 表示关键点的个数，Y_{xyc}表示真实值，$\hat{Y}_{xyc}$表示预测值。

由于在训练时对输入图像进行了 4 倍的下采样操作，这对最终的预测结果带来较大的误差，因此需要一个偏移量损失修正预测的结果。偏移量损失具体定义如下：

$$\mathrm{L}_{\mathrm{o}} = \frac{-\sum_{P}\left|\hat{B}_{\tilde{P}} - (\frac{P}{4} - \tilde{P})\right|}{N} \tag{2.13}$$

P 代表目标框中心点，4 表示下采样倍数，$\tilde{P}$表示预测目标的中心坐标，$\hat{B}_{\tilde{P}}$表示预测的中心点偏移量，$\frac{P}{4}$表示目标经过下采样之后的真实位置。

2.4.5 不同类型的卷积

(1)深度可分离卷积

深度可分离卷积[52]（Depthwise Separable Convolution）首先使用单个滤波器对每个输入通道进行深度卷积（Depthwise Convolution），然后通过逐点卷积（Pointwise Convolution）（1×1 卷积）将这些结果组合起来。通常卷积操作的输出通道数等于卷积核的数量，而深度可分离卷积对每个通道都只使用一个卷积核，所以单个通道在卷积操作之后的输出通道数也为 1。如果输入 Feature Map 的通道数为 N，对 N 个通道分别单独使用一个卷积核之后便得到 N 个通道为 1 的 Feature Map。再将这 N 个 Feature Map 按顺序拼接便得到一个通道为 N 的输出 Feature Map，其结构如图 2.13 所示。深度可分离卷积可以显著减少模型的参数数量和计算复杂度，因此这种卷积操作是设计轻量级深度学习模型的关键技术之一。

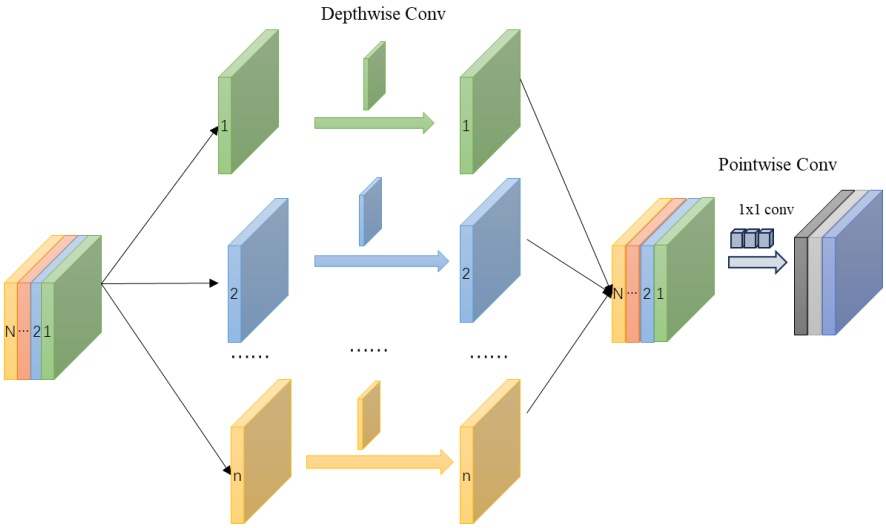


图 2.13 深度可分离卷积结构图[52]

(2)膨胀卷积/空洞卷积

通常在进行 Maxpooling 池化操作后，一些细节和小目标会丢失，在之后的上采样中无法还原这些信息，造成小目标检测准确率降低，而膨胀卷积/空洞卷积[53]（Dilated Convolution）通过引入额外的空间（空洞），也就是扩大卷积核的大小来增加滤波器的感受野，保持输入特征的宽高，而不增加参数数量或计算量，这使得模型能够捕捉更广阔的上下文信息，常用于图像分割和自然语言处理任务。其结构如图 2.14 所示，图中 5×5 的卷积结构在进行卷积运算时只有蓝色部分参与运算，而白色部分以 0 填充不参与卷积运算，数学表达式如 2.14 所示。

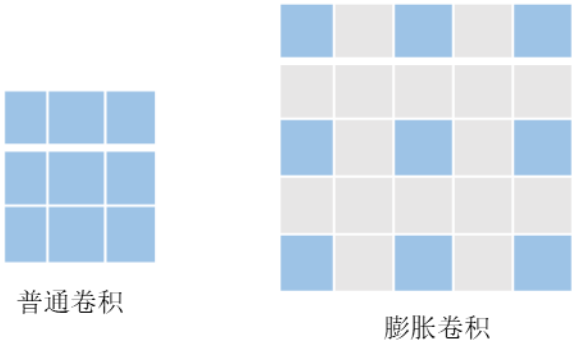


图 2.14 膨胀卷积示意图

$$K_c = S \times (K_o - 1) + 1 \tag{2.14}$$

其中K_c表示普通卷积核经过膨胀设计以后卷积核的尺寸，K_o表示普通卷积核尺寸，S 是一个超参数，表示膨胀因子。

(3) 可变形卷积

可变形卷积[54]（Deformable Convolution）是对传统卷积操作的一种扩展，允许网络学习卷积核采样位置的偏移量，这使得卷积核能够适应输入特征的几何变形，并且这种机制提高了卷积神经网络处理不规则几何变化的能力，尤其是对于图像中对象的扭曲、旋转等变化。在标准卷积中，卷积核的采样位置是固定的并且均匀分布在每个局部接受域内。相对地，可变形卷积为每个卷积核添加了额外的可学习的偏移参数，这些偏移参数决定了卷积核在输入特征图上的实际采样位置。换句话说，每个卷积核都可以自由地变形以更好地适应其覆盖的特征区域的形状和位置。可变形卷积使得模型能够捕获更加复杂的特征，提高了对图像变化的容忍度，能够提升网络性能。在二维情况下，对于给定的输入特征图 X 和卷积核 W，可变形卷积的输出 Y 在位置P_0的计算可表示为 2.15 所示：

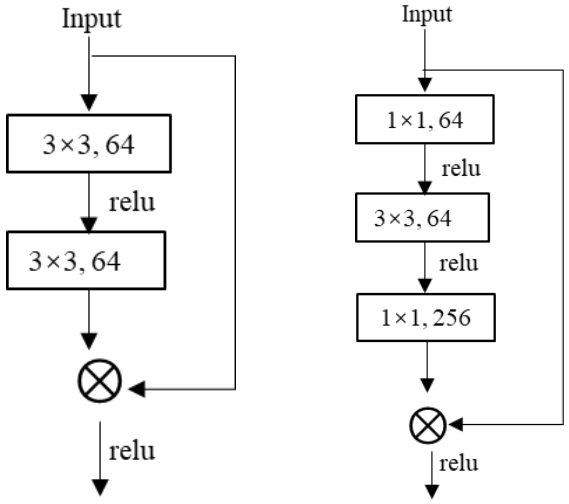
$$Y(p_0) = \sum_{P_n \in R} W(p_n) \cdot X(p_0 + p_n + \Delta p_n) \tag{2.15}$$

其中P_0是输出特征图中的位置，R是感受野区域的范围，P_n是卷积核W中相对于中心点的偏移位置，$W(P_n)$表示卷积核对应位置的权重，ΔP_n是学习到的偏移量，用于调整卷积核的采用位置，使之能够适应输入特征的局部变化，$X(p)$表示在位置 P 的输入特征图的值。

2.5 本领域常用模型

2.5.1ResNet 网络

ResNet（残差网络）由 He[55]等人在 2015 年提出，并在 ImageNet 大规模视觉识别挑战赛（ILSVRC）中取得了突破性的成绩，有效解决了深度神经网络中常见的梯度消失和梯度爆炸问题，使得网络能够达到前所未有的深度。ResNet 的核心概念是“残差学习”（Residual Learning）。传统的深度网络通过堆叠层来学习特征，但随着网络层的增加，训练这样的网络变得越来越困难，而且容易出现梯度消失或梯度爆炸的问题。残差学习的思路是通过引入“跳过连接”（Skip Connections）或“短路连接”（Shortcut Connections）实现让层去学习输入与输出之间的差异部分，而不是直接学习映射关系，从而在增加网络深度的同时保持了训练的稳定性。图 2.15 展示了两种不同残差结构，（a）通常用于 ResNet-50 等浅层结构中，（b）通常用于 ResNet-101 以及 151 等深层结构中。



（a）残差结构　　（b）残差结构

图 2.15 残差块

2.5.2Hourglass 网络

Hourglass（沙漏网络）[56]受沙漏形状设计的启发，旨在解决计算机视觉中的细粒度定位问题，如人体姿态估计和面部特征点识别，结构如图 2.16。其工作流程如下：首先进行初始的特征提取。接着进行重复多次的下采样，一般每次下采样通过步长为 2 的卷积或最大池化来实现。网络在每个下采样的阶段加入残差连接来增加特征提取的能力，并缓解深度模型训练中的梯度消失问题，确保网络深度的增加不会损害信息的有效传递。然后在到达沙漏的最底部之后，网络通常会有一些中间层，这些中间层可以是标准的卷积层或者残差模块，目的是进一步提取和处理特征。

从底部开始，通过使用转置卷积实现上采样，逐步恢复特征图的空间尺寸。在上采样的过程中，网络会通过跳跃连接将对应的下采样阶段的特征图与上采样的特征图进行融合，以保留多尺度的信息。该网络可以通过堆叠多个沙漏模块来不断细化特征表示，每个模块的输出都作为下一个模块的输入，以此来增强模型的学习能力和性能。Hourglass 模型通过其独特的多尺度特征整合能力和对称的堆叠结构，为处理复杂的视觉任务提供了一个强大且灵活的框架，使其在多个领域都表现出了卓越的性能。

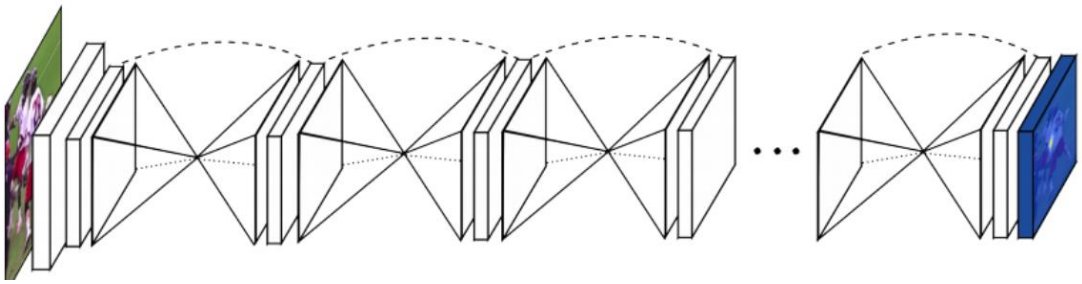


图 2.16 堆叠 Hourglass 网络结构[56]

2.6 本章总结

本章主要介绍了本文所使用的深度学习相关理论以及技术。首先介绍了早期神经网络，包括感知器、多层神经网络，并对其中涉及到激活函数、梯度下降法等技术进行了介绍；接着介绍了卷积神经网络，并重点介绍了网络中的卷积、池化、全连接等技术及其作用，并介绍了几种不同的卷积。最后对本章进行了总结。

第三章　基于注意力机制的头颅关键解剖点定位检测

3.1 引言

为了解决当前头颅侧位片测量领域存在的关键解剖点定位耗时长、准确率低、小样本下特征提取困难等问题，本文在目标检测算法CenterNet的基础上提出了基于注意力机制的头颅侧位片关键解剖特征点定位检测算法（CenterNet-Attention，CenterNetAt），并在公开数据集上进行了训练和验证。最终经过多次实验，发现融合注意力机制的模型能够提高解剖点的定位准确度和检测效率。

3.2 注意力机制

注意力机制（Attention mechanisms）是深度学习领域中的重要技术，其灵感来源于人类的视觉注意力系统。该系统能够在接收大量信息的同时，聚焦于某些关键信息。同样，在深度学习领域，这一概念被引入为一种能够使模型专注于输入数据的最重要部分的技术。它允许神经网络在处理输入数据时集中注意力于相关的部分，减少背景等信息的干扰，因此能够一定程度上提高模型的特征表达能力。同时，在计算能力有限情况下，注意力机制能够将有限的计算资源分配给更重要的任务，提高资源利用效率。自从Bahdanau等人[57]在2014年将其引入到序列到序列（Seq2Seq）模型中以来，注意力机制已经快速发展，并被广泛应用于各种任务中，显著提高了模型性能。注意力机制的核心思想是允许模型在处理输入数据时，对不同部分赋予不同的重要性。这种机制通常涉及三个关键元素：查询（Query）、键（Key）和值（Value）。查询是指当前正在处理的项，键指用于查找信息的标识符，值指与输入项相关联的实际信息内容。模型通过计算查询与键之间的匹配度来生成一个注意力权重（或得分），然后这些权重与值进行加权求和，得到一个输出结果，作为对查询的解答。当前常用注意力机制主要有挤压和激励网络（Squeeze-and-Excitation，SE）[58]，卷积块注意力机制（Convolutional Block Attention Module，CBAM）[59]，自注意力机制（Self-attention）[60]，十字交叉注意力机制（Criss-Cross Attention）[61]。

3.2.1 挤压和激励网络

挤压和激励网络（Squeeze-and-Excitation）主要用于通道维度的特征表示。它通过建模特征通道之间的相互依赖关系，动态调整通道的响应强度。主要工作流程如图 3.1 所示，首先输入包含 C 个通道的特征图，接着通过全局平均池化生成 1×1×C 的通道全局描述符 U，描述符可以看成是对全局信息的总结，即通过考虑整个空间范围内的特征来捕捉通道级的全局上下文信息；然后使用全连接层对生成的通道描述符 U 学习每个通道的注意力权重 X；最后将学习到的通道注意力权重与原始特征图逐通道相乘，得到带有权重的新特征图，最终强化了有用信息的捕捉能力。

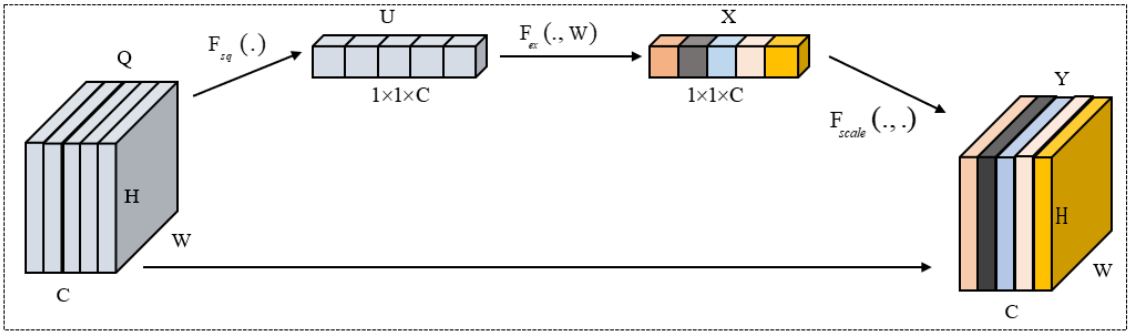


图 3.1 通道注意力机制[58]

3.2.2 卷积块注意力机制

卷积块注意力机制 (Convolutional Block Attention Module，CBAM)是一种轻量级注意力机制，包括通道注意力和空间注意力两个模块，如图 3.2 所示。

通道注意力机制主要用于设别输入特征图的哪些通道最为关键。简而言之，它通过评估不同通道的信息内容，决定哪些通道对当前任务更加重要。通过分析这些通道间的相互关系，通道注意力机制能够生成一个针对通道的注意力权重图，突出那些更重要的通道。具体来说，通道注意力机制的工作过程如下：给定一组输入特征图，其尺寸为高度 H、宽度 W 和通道数 C。首先，通过在通道维度上应用最大池化和平均池化操作，最大池化主要用于捕捉最显著的特征，而平均池化用于提取通道的平均信息，然后这两种操作生成的特征图尺寸都会被缩减为 1×1×C，即每个通道都被压缩成一个单一数值。接下来，这两个提取出的特征图被送入一个包含两层全连接层的多层感知机（MLP）中进行处理，得到经过感知机加工后的关键特征图和均值特征图。最终，将这两个特征图进行相加，并通过 Sigmoid 函数进行激活，以此生成最终的通道注意力特征图。该特征图通过赋予不同通道以不同的权重，引导模型关注于更加重要的通道特征，从而提升了模型的性能。

空间注意力块用于学习每个空间位置的重要性，关注数据的有效信息在“哪里”，以便网络可以在不同位置上分配不同的权重。它通过通道最大池化和全连接层生成空间注意力向量，并将其与原始特征图相乘，加权调整每个空间位置的特征响应，整个过程可以用式 3.1 和 3.2 表示。在实际的使用中，本文将 CBAM 的 Sigmoid 激活函数替换成 HSigmoid 以加快计算速度，降低计算复杂度，并且使用 HSigmoid 函数有助于减轻梯度消失问题。

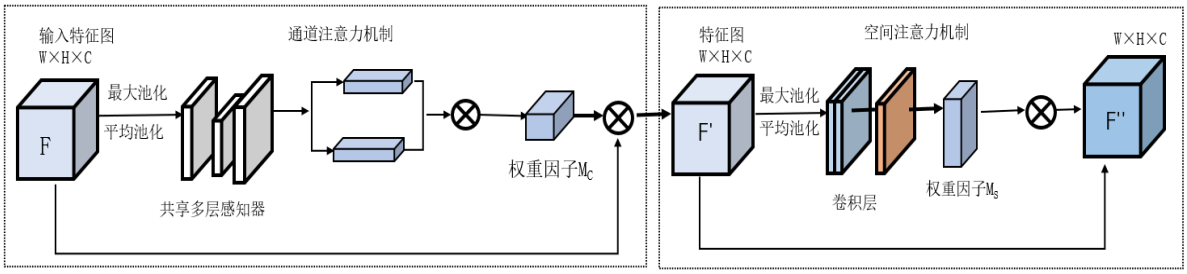


图 3.2 卷积块注意力机制

$$F' = M_C(F) \otimes F \tag{3.1}$$

$$F'' = M_S(F') \otimes F' \tag{3.2}$$

3.2.3 自注意力机制

自注意力机制（Self-Attention），也称为内部注意力（Intra-Attention）。它允许输入序列的每个位置直接建立联系并相互作用，从而能够捕捉序列内部的长距离依赖关系，因此也成为自然语言处理（NLP）和计算机视觉（CV）中任务中，获取序列内部动态关系的关键技术。自注意力机制通过计算序列中每个元素（如单词、像素等）对于序列中所有其他元素的注意力权重，来捕捉元素之间的依赖关系。这些注意力权重随后用于生成加权的输出，以此反映序列内部的复杂关系和结构。具体来说，自注意力机制涉及以下三个关键组成部分：查询（Query）：当前元素的表示，用于与键进行匹配。键（Key）：序列中所有元素的表示，用于与查询匹配。值（Value）：序列中所有元素的表示，其加权和将构成输出。每个元素的输出是所有值的加权和，其中权重由当前元素（查询）与序列中所有元素（键）之间的匹配程度决定。自注意力机制的工作流程如图 3.3 所示：首先输入序列被转换成三个不同的表示：查询（Query, Q）、键（Key, K）和值（Value, V），这些表示通过将输入序列的嵌入向量乘以三个可学习的权重矩阵得到。接着对于序列中的每个元素，通过

点积（Dot product）计算其查询向量与序列中所有键向量的相似度或匹配程度，然后使用 Softmax 函数对得分进行缩放和归一化，以得到注意力权重，这些权重表示了序列中每个位置对当前位置的重要性。最后使用上一步计算出的注意力权重，对所有的值（Value）向量进行加权求和。加权和后的结果是一个聚合了整个序列信息的输出向量，它在某种程度上反映了当前位置对序列其他部分的关注。对序列中每个位置都重复上述步骤，最终得到一个与输入序列同样长度的输出序列，其中每个元素都包含了整个序列的上下文信息。

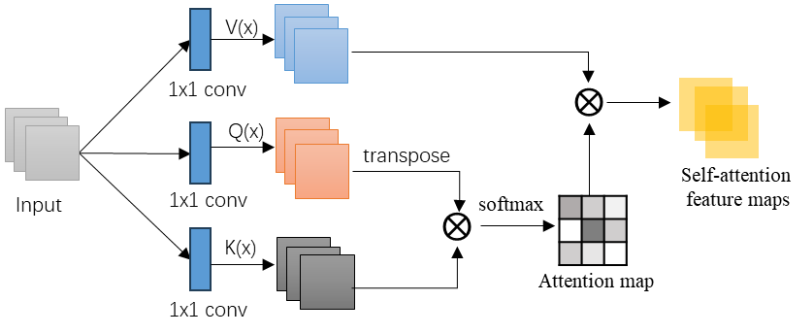


图 3.3 自注意力机制

3.2.4 十字交叉注意力机制

在自注意力中，每个序列元素（或特征位置）都会计算与序列中所有其他元素的关系，以确定每个元素对其他元素的注意力权重，这种机制使得每个元素都能够捕获到全序列的上下文信息，因此在处理长序列或大规模二维数据时可能会面临较高的计算复杂度，需要较大的内存和计算资源。而十字交叉注意力机制（Criss-Cross Attention）通过两个独立的扫描过程（垂直和水平）来聚集每个元素在其所在行和列的全局信息。十字交叉注意力机制允许每个位置的元素聚焦于其所在行和列的其他元素，这种机制允许模型在保持计算效率的同时，捕获图像的全局依赖关系。通过限制注意力的范围到每个元素的行和列，十字交叉注意力能够在一定程度上减少计算量，尤其是在处理具有大尺寸的二维数据时。其工作原理如 3.4 下：输入特征图首先通过一个卷积层，生成用于注意力机制的查询（Query）、键（Key）和值（Value）。接着进行垂直和水平扫描，在垂直扫描过程中，对于特征图上的每个像素，模型计算其与同一列中所有像素的注意力权重。在水平扫描过程中，模型对每个像素计算其与同一行中所有像素的注意力权重。通过这两次扫描，每个像素点都能够获得其所在行和列的全局信息。然后模型将这些信息聚合到每个位置，以获取

全局上下文信息。最后，原始将输入特征和通过十字交叉注意力得到的特征进行融合（通常是相加或拼接），以生成最终的带有注意力权重的特征图。

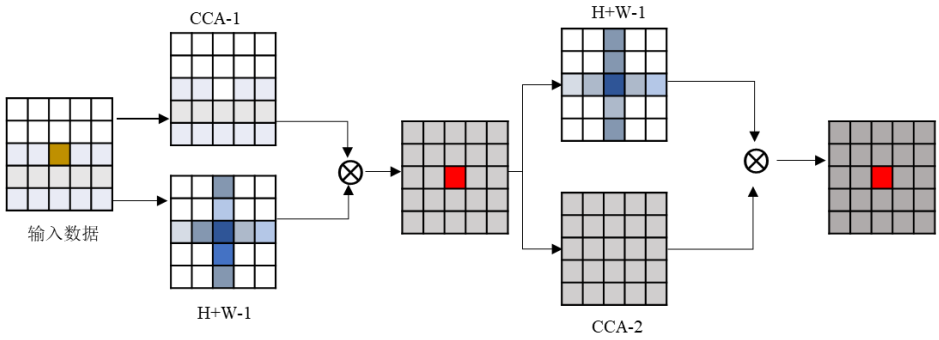


图 3.4 十字交叉注意力机制

3.3 改进的 CenterNet 关键点定位检测网络

整体网络：本文采用端到端的目标检测算法 CenterNet 作为研究的基准网络，其核心思想是将目标检测任务转化成关键点（目标中心点）检测任务。与传统基于锚框（Anchor-base）的目标检测方法（如 Faster R-CNN[62]、SSD[63]等）不同（基于锚框的算法通常是指算法首先需要生成大量预定义的、不同尺寸和形状的锚框，然后通过非极大值抑制（Non-Maximum Suppression，NMS）等算法筛选这些锚框，找出最佳匹配目标对象的框，然后再分类等），CenterNet 直接预测目标的位置或者大小， 避免了锚框的使用（Anchor-free），简化了网络结构，提高了检测效率。重要的是，CenterNet 不仅能够在简化的网络结构中保持较高的检测精度，还具有较快的推理速度，非常适用于头影关键解剖特征点定位检测这种实时应用场景。CenterNet 工作流程如图 3.5 所示：输入图像首先经过预处理（缩放，归一化等）以满足网络对于输入数据大小的要求（1024×1024），将经过预处理后的图像送入专门用于特征提取的骨干网络（Backbone Network）中进行特征提取，经过骨干网络中一系列的卷积、池化、下采样等操作后产生特征图。利用特征提取得到的特征图，模型通过一个卷积头输出目标的中心点热图。该热图的每个像素值代表相应位置成为目标中心点的概率。最终通过查找热图的局部最大值来确定目标中心点的位置。由于在特征提取的过程中使用了下采样技术使得热图的分辨率通常低于原始图像，导致直接从热图的局部最大值点映射回原始图像时可能会存在定位误差，因此模型还会对热图上的每一个局部最大值点，输出一个 x 方向和 y 方向的偏移量，用于修正中心点

的位置，通过偏移量纠正下采样导致的精度损失，目标的定位更加精确。而在目标检测等任务中，网络还会同步回归用于调整目标框位置的损失。

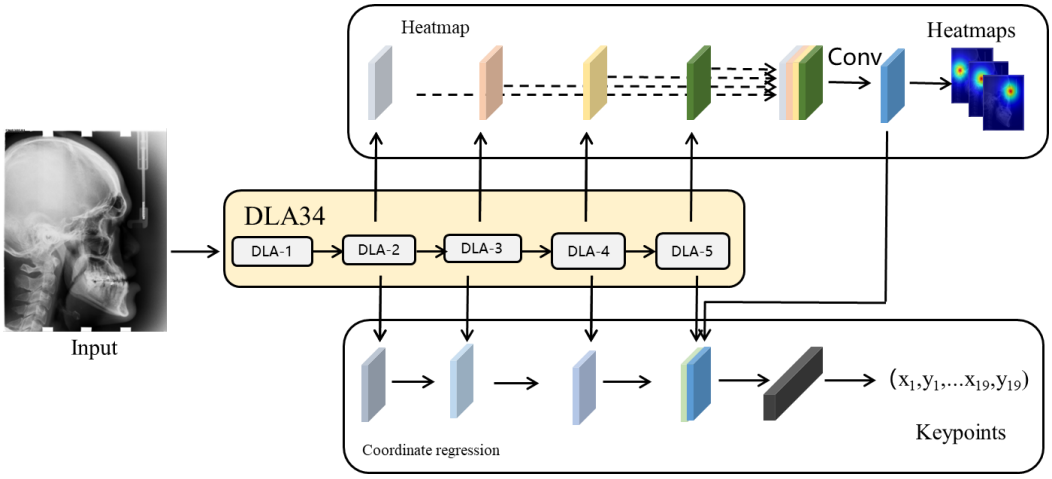


图 3.5 CenterNet 结构图

骨干网络：本文使用的主干网络为深度聚合网络 （Deep Layer Aggregation，DLA34），其结构如图 3.6 所示。DLA 主要由 Conv Blocks（基础卷积块）、Aggregation Node（聚合节点）、Iterative Deep Aggregation（迭代深度聚合）、Hierarchical Deep Aggregation（分层深度聚合）4 个部分组成。Conv Blocks 是网络中的基础块，并通过逐层堆叠的方式构成网络的基本结构，每个 Conv Blocks 卷积块负责从输入数据中提取不同层次的特征；Aggregation Node 聚合节点是 DLA 网络中用于实现特征聚合的基本单元，每个聚合节点采用相加（Add）或拼接（Concat）操作将不同路径或不同层的特征进行融合；Iterative Deep Aggregation 主要是通过迭代将同一尺度下的深层特征与前面层的特征结合，逐步构建更深、更丰富的特征表示，强化网络在同一尺度内特征的深度和复杂性，提高模型捕获更细粒度信息的能力。Hierarchical Deep Aggregation 主要作用是将不同尺度的特征映射和融合到同一尺寸下，实现跨尺度的信息交流，使得模型能够同时利用高层次的语义信息和低层次的细节信息，提高模型对上下文和局部细节的理解能力。与 ResNet 等跳跃连接网络只是进行简单叠加不同，DLA 网络通过设计这种多尺度特征聚合的方式使得网络能够更好的融合不同层次和尺度的特征信息，最终使模型有更高的精度的同时有更少的参数。