计算机科学
COMPUTER SCIENCE

基于类注意力的眼睛凝视估计网络

徐金龙, 董明瑞, 李颖颖, 刘艳青, 韩林

引用本文

徐金龙, 董明瑞, 李颖颖, 刘艳青, 韩林. 基于类注意力的眼睛凝视估计网络[J]. 计算机科学, 2024, 51(10): 295-301.

XU Jinlong, DONG Mingrui, LI Yingying, LIU Yanqing, HAN Lin. Eye Gaze Estimation Network Based on Class Attention [J]. Computer Science, 2024, 51(10): 295-301.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230900094

基于类注意力的眼睛凝视估计网络

徐金龙[1,3]　董明瑞[1,2]　李颖颖[1,3]　刘艳青[1]　韩　林[1]
1 国家超级计算郑州中心　郑州 450000
2 郑州大学计算机与人工智能学院　郑州 450000
3 信息工程大学　郑州 450000
(longkaizh@163.com)

摘　要　近年来,眼睛凝视估计引起广泛关注。基于RGB外观的凝视估计方法使用普通摄像机和深度学习来进行凝视估计,避免了像商用眼动仪一样使用昂贵的红外设备,为更准确和成本更低的眼睛凝视估计提供了可能。然而,RGB外观图像中包含如光照强度、肤色等多种与凝视无关的特征,这些无关特征会在深度学习回归的过程中产生干扰,进而影响凝视估计的精度。针对以上问题,提出了一种名为类注意力网络(CA-Net)的新架构,它包含通道、尺度、眼睛3种不同的类注意力模块,通过这些类注意力模块可以提取和融合不同种类的注意力编码,从而降低与凝视无关特征所占的权重。在GazeCapture数据集上的大量实验表明,在基于RGB外观的凝视估计方法中,相比现有的最先进方法,CA-Net在手机和平板上分别能够提高约0.6%和7.4%的凝视估计精度。

关键词:类注意力;轻压缩激励;自注意力;多尺度;眼睛凝视估计
中图分类号　TP183

Eye Gaze Estimation Network Based on Class Attention

XU Jinlong[1,3],DONG Mingrui[1,2],LI Yingying[1,3],LIU Yanqing[1] and HAN Lin[1]
1 National Supercomputing Center in Zhengzhou,Zhengzhou 450000,China
2 School of Computer and Artificial Intelligence,Zhengzhou University,Zhengzhou 450000,China
3 Information Engineering University,Zhengzhou 450000,China

Abstract　In recent years,eye gaze estimation has attracted widespread attention. The gaze estimation method based on RGB appearance uses ordinary cameras and deep learning for gaze estimation,avoiding the use of expensive infrared devices like commercial eye trackers,providing the possibility for more accurate and cost-effective eye gaze estimation. However,due to the presence of various features unrelated to gaze,such as lighting intensity and skin color,in RGB appearance images,these irrelevant features can cause interference in the deep learning regression process,thereby affecting the accuracy of gaze estimation. In response to the above issues,this paper proposes a new architecture called class attention network(CA-Net),which includes three different class attention modules:channel,scale,and eye. Through these class attention modules,different types of attention encoding can be extracted and fused,thereby reducing the weight of gaze independent features. Extensive experiments on the GazeCapture dataset show that,compared to the state-of-the-art method,CA-Net can improve gaze estimation accuracy by approximately 0.6% and 7.4% on mobile phones and tablets,respectively,in RGB based gaze estimation methods.

Keywords　Class attention,Light squeeze-and-excitation,Self-attention,Multiscale,Eye gaze estimation

1　引言

凝视估计作为一项计算机视觉任务,旨在利用计算机来预测人眼在观看图像或视频时的凝视点。这一技术在实际中有着广泛的应用,如检测驾驶员的疲劳程度[1]、分析用户在注意力分布和偏好方面的行为(如广告和网页浏览)[2]和无障碍人机交互[3]。随着计算机算力和深度学习的迅猛发展,与传统的眼睛凝视估计方法(如电眼描记术、螺旋搜索线圈等[4])相比,基于深度学习的眼睛凝视估计方法具备高准确性和实时性的优势,并且无需对被试者进行任何侵入性的物理操作。因此,本文基于深度学习的方法对眼睛凝视估计算法开展研究。

到稿日期:2023-09-18　返修日期:2024-01-16
基金项目:2022年河南省重大科技专项(221100210600);22求是科研启动(自)(32213247);2023年度河南省科技攻关专项(232102210185)
This work was supported by the 2022 Henan Province Major Science and Technology Special Project(2211002110600),22 Qiushi Research Initiation(Natural Science)(32213247) and 2023 Henan Province Science and Technology Research Special Project(232102210185).
通信作者:韩林(hanlin@zzu.edu.cn)

现有的凝视估计算法大致可以分为基于模型的方法和基于外观的方法。基于模型的方法通过建立眼睛模型来预测注视点，然而，由于每个人眼睛形态的独特性，这种方法需要为每个人进行眼睛模型的个性化校准；而基于外观的方法则借助于深度学习从外观图片映射得到注视点坐标，深度神经网络强大的学习能力配合大量公开的凝视估计数据集使得基于外观的方法可以直接为不同的人进行凝视估计。

商用眼动仪使用的方法属于基于外观的方法，部分商用眼动仪声称在实际场景中能够将凝视角度误差控制在0.5°～1°之间。然而，商用眼动仪需要使用专用设备（如红外相机）来捕捉外观图像，这不利于将凝视估计应用于更广泛的群体。因此，当前大多数基于外观的研究集中在使用普通相机（如网络摄像头、手机和平板电脑的前置摄像头）拍摄的RGB图像来进行凝视估计上，而本研究也是使用RGB图像进行基于外观的注视点坐标的预测。

由于与凝视相关的信息主要集中在眼睛部位，因此大多数基于外观的眼睛凝视估计算法通常将截取的眼睛图像和截取眼框的位置坐标作为输入。此外，一些方法还会添加脸部图像作为输入以进行头部姿态估计[5]或指导双眼特征的融合[6]，从而提高算法的准确度。从输入数据的角度来看，眼睛凝视估计的难点在于如何从眼睛图像中提取和融合凝视相关且适用于更多人和场景的鲁棒特征。为解决这一问题，本文提出了类注意力网络。如图1所示，类注意力网络中不同的类注意力模块可以根据不同的维度（通道、尺度、眼睛）将眼睛特征划分为不同的种类，并为不同种类的眼睛特征添加不同的注意力权重（类注意力编码），从而降低与眼睛凝视无关的眼睛特征所占的权重，进而提高眼睛凝视估计的精准度。此外，本文受压缩激励网络SENet（Squeeze-and-Excitation Network）[7]启发，提出了轻压缩激励模块。该模块通过轻压缩代替SENet的全局平均池化操作，保留了更多的特征图信息，使网络获取了更精准的通道注意力。

1）提出了类注意力网络，用于眼睛凝视估计任务。该网络在特征融合过程中按照类型信息自适应地提取和融合注意力，更好地引导网络去关注与眼睛凝视相关的特征。

2）提出了轻压缩激励模块。该模块通过轻压缩和多头注意力提供了更精准的通道注意力。

3）在GazeCapture数据集上使用小分辨率眼睛图像（30×30像素）获得了目前最优结果，约2.13 cm（平板）和1.61 cm（手机）的误差精度。

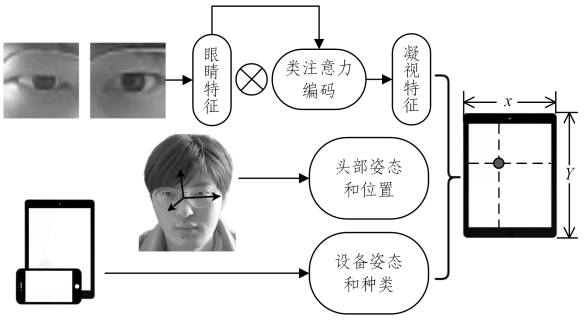


图1 所提方法的说明

Fig. 1 Description of the proposed method

本文第2章将介绍眼睛凝视估计的相关研究；第3章将详细介绍提出的网络结构；第4章将展示在GazeCapture数据集上进行的消融研究、和其他先进工作的对比，以及对类注意力编码含义的分析；最后总结全文并展望未来。

2 相关工作

凝视估计一直是备受关注的研究领域，研究者们在这个领域提出了许多方法。这些方法按照求解方式的不同基本可以分为基于模型的方法和基于外观的方法两类。

2.1 基于模型的方法

基于模型的方法旨在利用眼睛特征，如角膜反射、瞳孔中心和虹膜轮廓等，来建立与眼睛运动和姿态相关的模型，从而实现对视线方向和凝视点的估计[8]。在这方面，Sun等[9]提出使用RGB-D摄像机来收集参与者的面部和瞳孔中心的3D信息，然后通过3D几何眼睛模型进行凝视估计。他们还考虑了头部移动和视轴与光轴的偏差，以减少凝视误差。与之不同，Dementhon等[10]没有使用RGB-D摄像机，而是提出了一种通过拟合2D人脸关键点到3D人脸模型来确定人脸3D位置的方法。此外，Wang等[11]提出了一种基于单个网络摄像头的可变形的眼脸模型，该模型可以通过统一校准算法恢复单个3D眼-脸模型和个人眼睛参数，进一步提高了凝视估计的鲁棒性和准确性。基于模型的方法大多受制于用户-相机距离以及专业设备的可用性，如红外摄像机和RGB-D摄像机。因此，基于模型的方法尽管在受控实验室环境下表现准确，但在无约束环境下可靠性较低[6]。

2.2 基于外观的方法

基于外观的方法可以从外观图像直接映射得到凝视点的估计坐标。根据拍摄图像设备的不同，基于外观的方法又可分为基于红外外观的方法和基于RGB外观的方法两类。商用眼动跟踪器，如Tobii X系列，采用的是基于红外外观的方法，原理是通过捕获角膜和视网膜反射的红外光来获取眼睛运动特征，从而进行凝视估计。然而，由于红外相机具有非通用性，目前基于外观的研究大多集中于基于RGB外观的方法上。基于RGB外观的方法采用深度学习技术以增强特征检测器，提高对光照变化的鲁棒性。此外，由于大量RGB图像的凝视估计数据集已被公开，基于普通RGB相机的外观方法在现实世界的非受控环境中表现出了较高的准确性和鲁棒性，因此受到广泛关注。其中，GazeCapture数据集由Krafka等[12]收集并发布，包含1 400多名参与者和240多万个样本，该数据集引入了一种名为iTracker的凝视估计新框架。iTracker模型使用左眼图像、右眼图像、人脸图像以及人脸位置作为输入，经不同的神经网络提取特征后融合回归，最终得到凝视点位置。He等[13]认为人脸图像和人脸位置之间存在信息冗余，将这两个输入替换为眼框坐标并提出了SAGE模型，这种替换在保持精度的同时提高了推理速度，在Google Pixel 2手机上可达到每帧10毫秒的处理速度。Guo等[14]则在网络训练中引入蒸馏和剪枝技术以防止过拟合，取得了更为鲁棒的结果。Bao等[6]认为双眼特征融合过程被忽略，提出了AFF-Net，通过自适应特征融合模块使用人脸信息来指导双眼特征融合，从而获得更精准的估计结果。Athavale

等[15]使用 ResNet 和 Inception 架构实现了一个凝视估计网络，并提出了一种集成校准网络来针对特定主题进行预测，以较小的模型参数量实现了高性能。然而这些研究均没有关注到特征种类信息对最终凝视结果的影响，在眼睛特征融合的过程中，不同类的特征会对最终的凝视估计做出不同的贡献，而忽略这种贡献的差异性最终会导致凝视估计精准度下降。本文提出的 CA-Net 通过依次为不同通道、不同尺度、不同眼睛的特征添加注意力来指示不同类的特征组的重要性，从而增加了与凝视相关特征所占的权重，提升了凝视估计的精准度。

3 网络结构

3.1 类注意力网络

类注意力网络(CA-Net)的结构如图 2 所示。CA-Net 的输入包含 3 个部分，分别为左眼和翻转的右眼图像、头部姿态的 3D 向量和 3D 位置坐标、设备姿态和种类的 one-hot 编码。

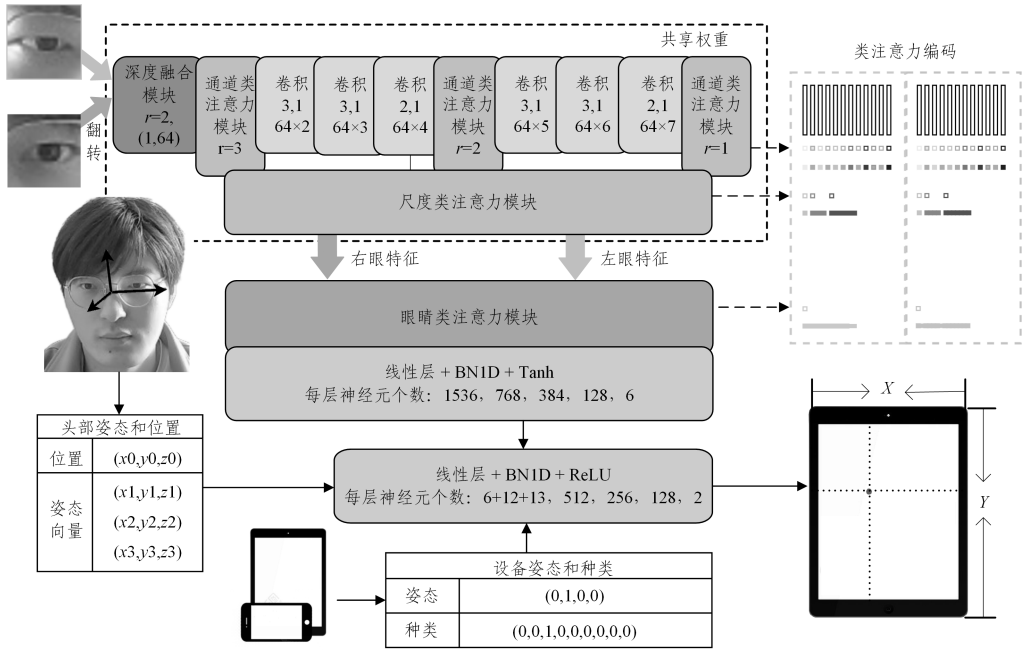


图 2 CA-Net 的结构。

Fig. 2 Structure of CA-Net

左眼和翻转的右眼图像分别被输入主干网络中进行特征提取，并通过类注意力模块进行特征增强，接着将增强的眼睛特征经线性层压缩后与头部姿态向量、位置坐标以及设备姿态和种类的 one-hot 编码拼接，拼接结果再经线性层回归得到最终的凝视点坐标。主干网络由卷积块堆叠而成，输出多尺度的眼睛特征。类注意力模块与自注意力模块[16]类似，可以自适应地提取并融合类注意力编码，进而降低与凝视无关特征所占的权重。此外，考虑到左右眼结构的相似性，在左眼特征和右眼特征被眼睛类注意力模块融合之前的网络均共享权重。

3.2 类注意力模块

类注意力模块通过所有类的特征来确定不同类的特征组应分配的权重，并通过乘法将种类信息融合到特征中，以便网络能更关注与凝视估计相关的特征。本文根据通道、尺度和眼睛 3 种不同的划分方式，分别设计了通道类注意力模块、尺度类注意力模块和眼睛类注意力模块。这些模块的工作原理均可由式(1)表示：

$$\begin{aligned} &W_1, W_2, \cdots, W_n = \phi(\Gamma_{\text{concat}}(f_{\text{in}_1}, f_{\text{in}_2}, \cdots, f_{\text{in}_n})) \\ &f_{\text{out}} = \Gamma_{\text{concat}}(f_{\text{in}_1} * W_1, f_{\text{in}_2} * W_2, \cdots, f_{\text{in}_n} * W_n) \end{aligned} \tag{1}$$

其中，f_{in}表示类注意力模块的输入，其在通道类注意力模块中代表前面卷积网络输出的每一个通道特征，在尺度类注意力模块中代表不同深度卷积层输出的不同尺度的特征经由通道类注意力模块增强后输出的特征组，在眼睛类注意力模块中代表左右眼睛分别对应的所有尺度类注意力模块输出拼接后的特征组。f_{in_n}表示第 n 个的特征组。$\Gamma_{\text{concat}}(.)$表示拼接操作，将所有特征组连接在一起。在通道类注意力模块中，由于通道特征天然连接在一起，因此该步骤可以省略。$\phi(\cdot)$表示计算类注意力权重的方法，其在通道类注意力模块中代表轻压缩激励操作，在尺度类注意力模块和眼睛类注意力模块中代表线性层。W_n 表示按照某一划分方式(通道、尺度和眼睛)进行分类时第 n 个类的特征组应分配的注意力权重，W_1，W_2，…，W_n 组合在一起即表示该划分方式所对应的类注意力编码。原始输入的特征组与相应的注意力权重分别相乘，然后通过 $\Gamma_{\text{concat}}(\cdot)$操作拼接为一个整体(通道类注意力模块省略该操作)，从而获得最终的输出 f_{out}。

3.2.1 通道类注意力模块

通道注意力最早由 SENet 引入，用于指示特征图中每个通道的重要程度。不同通道的特征对凝视估计有不同的重要性。例如，眼睛图片的光照特征与凝视估计无关，而眼睛的张开程度、瞳孔相对眼眶位置等与凝视估计密切相关。因此，对不同通道的眼睛特征使用通道注意力是必要的。然而，SENet 使用全局平均池化层将通道特征压缩到 1×1 大小来代表每个特征图的信息，过度的压缩使得特征的相关性表达能力受限，因此，本文提出了轻压缩激励模块，将特征图轻压缩至 5×5，保留了更多的相关性信息，并且设计了自注意力

模块来计算二维特征图之间的注意力，以更精确地指示每个通道特征的重要性。

如图 3 所示，通道类注意力模块主要由轻压缩激励模块构成。相对于传统的压缩激励(SE)模块，轻压缩激励模块采用深度融合模块替代平均池化操作，以保留高维特征图信息。深度融合模块首先使用 PixelUnshuffle 算子(PixelShuffle[17]的逆算子)将$[N,C,H,W]$的张量重新排列成$[N,r^2C,H/r,W/r]$的形式，随后使用 1×1 的深度卷积进行降维，其中 r 表示减小空间分辨率的缩放因子，即将特征图缩小为原来的百分比。通过这一轻压缩操作，不仅在维持特征图空间偏置的同时保留了更多的高维特征信息，而且由于使用的是 1×1 的卷积，模型推理的速度会比直接使用步长为 r 的 $r\times r$ 卷积更快。此外，为了适配二维特征图之间的注意力求解，轻压缩激励模块还使用了多头自注意力模块来替代压缩激励模块中的线性层。与 Transformer[16]中的操作类似，轻压缩激励模块将二维特征图与位置编码相加，其中使用的位置编码是类似于 Bert[18]中的可学习位置编码，且所有通道共享相同的位置编码。然后，将特征图展平为一维特征向量，并通过多头自注意力模块来计算特征向量的注意力。最后，通过线性层将注意力向量压缩为注意力值，从而解决了计算二维特征之间注意力的问题。轻压缩激励模块的通道注意力输出一方面直接与原始特征图相乘，输入图 2 中后续的卷积层；另一方面，与轻压缩特征图相乘，并通过平均池化操作，用作图 2 中尺度类注意力模块的输入。

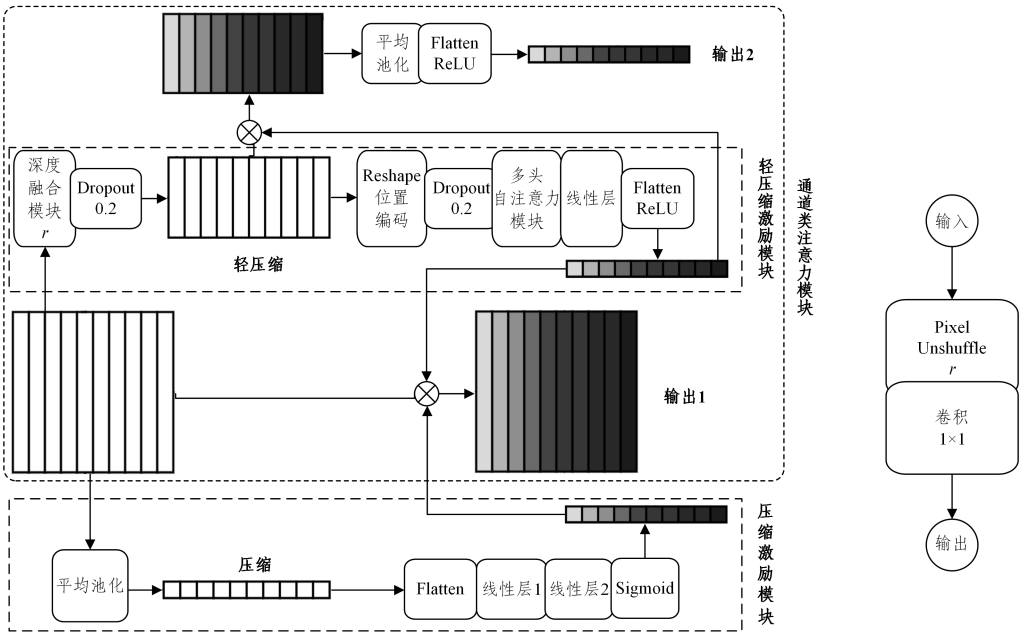


图 3 通道类注意力模块的结构

Fig. 3 Structure of channel class attention modules

3.2.2 尺度类注意力模块和眼睛类注意力模块

不同尺度的特征包含不同类型的信息。浅层特征包含具体的空间信息，而深层特征包含更抽象的高级信息。因此，若为这些不同尺度的特征赋予同样的权重会让神经网络难以区分这些不同类型的信息，进而影响凝视位置的预测。

再者，人类在进行视觉观察时，双眼中往往有一只被称为主视眼，它在视觉处理中占据主导地位，大脑更倾向于依赖主视眼的图像信息来分析和定位物体[19]。这一现象揭示了双眼在协同凝视过程中并非同等重要，而是有所侧重。

鉴于上述观察，本文分别设计了尺度类注意力模块和眼睛类注意力模块，来为不同尺度、不同眼睛的特征施加合适的权重。值得注意的是，由于输入的特征是一维的，因此这里并没有采用轻压缩激励模块来计算这两类注意力模块的注意力，而是采用了类似 SENet 的结构。

如图 4 所示，尺度类注意力模块和眼睛类注意力模块的结构相似。首先将不同类别的眼睛特征拼接成一个整体，然后使用线性层来计算每个类别的注意力权重，最后将这些权重与原始输入对应相乘后拼接作为输出。尺度类注意力模块的输入为 3 个来自不同尺度的特征组，输出加强的多尺度眼睛特征。眼睛类注意力模块的两个输入是来自两只眼睛的特征，输出为加强的双眼特征。

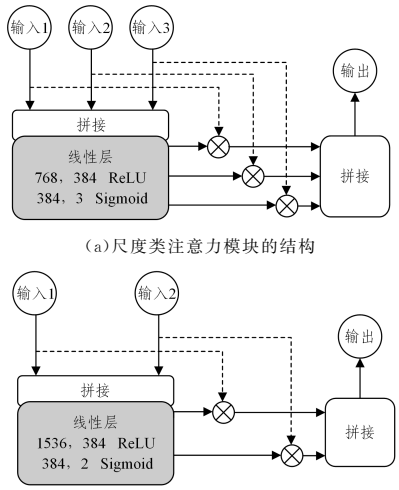


图 4 尺度类注意力模块和眼睛类注意力模块的结构

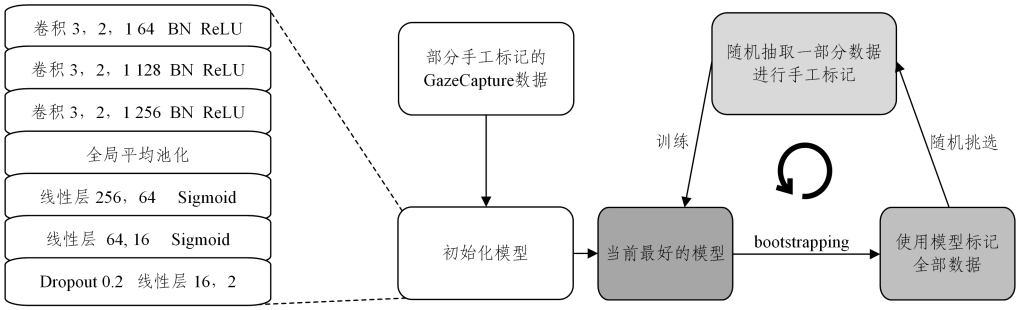
Fig. 4 Structure of scale class attention modules and eye type attention modules

4 实验

4.1 数据集和数据预处理

实验基于著名的 2D 凝视标签数据集 GazeCapture 开展。该数据集包含 1 474 名参与者，总计约 2 445 504 帧图像，通过 iPhone 和 iPad 采集而来。其中有效帧为 1 490 959 帧（即同时检测到眼睛和人脸），分为 1 251 983 张训练图像、59 480 张验证图像和 179 496 张测试图像。GazeCapture 为每帧图像提供了注视点标签，标签值是以相机为原点的坐标系下凝视位置的像素的物理位置，分别以向右和向下为正方向，给出物理坐标。此外，GazeCapture 还提供了每帧图像对应的设备姿态和种类信息，设备姿态根据 home 键朝向分为向上、向下、向左和向右 4 类，iPad 有 7 个设备种类，iPhone 有 8 个设备种类。

由于未睁眼的凝视数据会误导网络，使网络无法学到正确的凝视估计方法[20]，因此本文设计了一个简单的判别网络 Valid Net，用于筛选出睁眼的可用图像。Valid Net 的结构如图 5(a)所示，其核心结构由一系列精心构建的卷积层堆叠构成，这些卷积层可以高效地从眼部图像中提取关键特征信息。随后，这些特征信息通过多层线性层的进一步处理，最终生成分类结果。由于 GazeCapture 是一个非常庞大的数据集，手动标记整个数据集将会非常耗时。因此，本文采用了图 5(b)所示的 bootstrap 的数据标记法来标记整个数据集，并同时训练了 Valid Net。具体来说，该方法首先使用少量手工标注数据训练 Valid Net，赋予其初步分类能力。随后，选择表现最佳的模型自动标注剩余未标注数据。为提升数据质量，随机选取部分数据进行手工复核并更新标注。之后，利用更新后的全数据集重新训练 Valid Net，并迭代选择最优模型进行下一轮自动标注。该循环过程持续进行，直至手工复核中发现的错误标注大幅减少，以确保数据准确性和模型稳定性。未睁眼图像和睁眼图像的标签分别为 0 和 1，Valid Net 为每一张数据输出 0～1 的浮点数标签。使用 0.65 的阈值筛选整个数据集，筛选后的数据集包括 976 182 张训练图像、48 557 张验证图像和 150 352 张测试图像。此外，本文使用 mediapipe 的人脸地标网络[21]标记了每帧图像中参与者人脸的 468 个按图像尺度归一化的 3D 关键点坐标，利用眼部的关键点坐标抠取了眼睛图片，并使用 6 号关键点坐标作为头部位置，通过对 6 号、127 号和 365 号关键点坐标的计算，获得 3 个正交的单位方向向量，表示头部姿态。按照设备屏幕尺度，iPhone 和 iPad 设备一共被分为 9 个种类，与设备姿态一样，被映射为 one-hot 编码，以考虑设备差异性。



(a) Vaild Net 结构　　(b) Bootstrap 的数据标记和训练方法

图 5　Valid Net 的结构以及 bootstrap 的数据标记和训练方法

Fig. 5　Structure of Valid Net and data labeling and training methods of bootstrap

4.2 训练设置

与其他 GazeCapture 上的研究一样，本文遵循了 GazeCapture 数据集预设的训练数据集和测试数据集划分。此外，眼睛图像被统一调整为 30×30 的灰色图像，并对训练集图像使用了一组数据增强方法，即随机亮度转换、随机对比度转换和随机透视变换，以提高网络的鲁棒性。采用欧氏距离作为损失函数，其计算式如下：

$$loss=\sqrt{\sum_{i=1}^{n}(pre_i-lable_i)^2} \tag{2}$$

其中，pre 和 $lable$ 分别表示网络的预测值和标签值，它们在对应位置 i 上计算差值。在本文的 2D 凝视估计中，n 为 2，所求损失值等于物理凝视距离的误差值。CA-Net 在 GazeCapture 数据集上进行了 11 轮训练，前 5 轮的学习率为 0.000 5，接下来的 6 轮学习率为 0.000 1。使用 Adam 作为优化器，设置权重衰减为 0.000 1，批量大小设为 250。网络使用 Paddle 编写，并由 Paddle 完成网络权重的初始化。

4.3 实现细节

用于多尺度眼睛特征提取的主干网络的深度融合模块设置了缩放因子 r 为 2，并且其中 1×1 卷积核的数目为 64。此外，主干网络中含有 6 个卷积层，卷积核数目、卷积核大小、卷积步长、填充数目分别为[(64×2,3,1,0),(64×3,3,1,0),(64×4,2,1,0),(64×5,3,1,0),(64×6,3,1,0),(64×7,2,1,0)]，同时使用 BN 层和 ReLU 函数作为卷积层的标准化层和激活函数。通道类注意力模块中使用多头注意力模块（均为 5 头自注意力模块）求取高维特征的相关性，并且设置 Dropout 为 0.2。多尺度特征和通道类注意力编码融合后被压缩为一维，其大小从左向右依次为[N,64]，[N,64×4]和[N,64×7]，其中 N 为批量大小。多尺度特征分别与尺度类注意力权重和眼睛类注意力权重相融合，然后经过一系列线性层回归，得到形状为[N,6]的眼睛特征。这些线性层均经过 BN 层标准化，并使用 Tanh 函数激活。最后，眼睛特征与头部姿态向量和位置坐标、设备姿态和种类的 one-hot 编码拼接，通过多层线性层回归最终得到凝视点，这里的线性层（除最后一层外）均经过 BN 层标准化且由 ReLU 函数进行激活。

4.4 消融实验

在和其他的凝视估计模型进行对比之前，本文先评估了CA-Net架构的有效性。以CA-Net为基准，先后测试了在移除其中的通道类注意力模块、尺度类注意力模块、眼睛类注意力模块或设备种类信息时网络的精度，以及将通道类注意力模块中的轻压缩激励模块替换为压缩激励(SE)模块时网络的精度。结果如表1所列。

表1 消融实验的结果对比。

Table 1 Comparison of results from ablation experiments

(cm)

方法	手机凝视误差	平板凝视误差
CA-Net(基准)	1.58	2.13
缺少通道类注意力模块	1.58	2.19
缺少尺度类注意力模块	1.60	2.16
缺少眼睛类注意力模块	1.59	2.18
缺少设备种类信息	1.70	2.50
CA-Net(SE)	1.62	2.19

如表1所列，当移除各类注意力模块时，网络的精度均有所下降，且在手机上和平板上的下降幅度并不相同。在小尺度设备(手机)上进行凝视估计时，尺度类注意力模块对模型精度的增益最大，可以达到约1.3%，通道类注意力模块对模型精度的增益并不明显。相反，在大尺度设备(平板)上进行凝视估计时，通道类注意力模块可以达到约2.8%的增益，反而最大。尺度类注意力模块的增益最小，约为1.4%。眼睛类注意力模块始终为网络提供稳定的增益，约为0.6%(手机)和2.3%(平板)。

在其他基于GazeCapture的研究中，并没有工作将设备种类信息作为网络的输入。然而，在本文的实验中发现设备种类这一信息对凝视估计的精度产生了显著影响。如表1所列，以CA-Net为基线，若剔除设备种类信息，网络在手机和平板上将分别增加约7.6%和17.4%的平均误差。GazeCapture使用物理距离记录凝视坐标，所有设备均相对于设备摄像头构建坐标系，凝视坐标与设备类型不相关。然而，实验结果却证明了设备种类信息对凝视估计非常重要。本文推断设备种类可能为不同设备的凝视点提供了不同的屏幕尺度信息，而这一信息使得预测结果能够处于屏幕尺度内，从而提高了凝视估计的精度。

将轻压缩激励模块替换为SE模块意味着每个特征图信息的特征大小由5×5降至1×1，过度的压缩导致计算注意力时可用信息的细节丢失，进而导致注意力精度和凝视估计精度的下降。从实验结果可知，使用轻压缩激励模块的CA-Net(基准)的精度会比使用SE模块的CA-Net高约2%(手机)和3%(平板)。而如果直接使用原始的特征图来计算注意力，则过长的特征又会导致自注意力的计算和内存成本疯狂飙升。因此，本文提出的轻压缩激励模块是在综合考虑成本和性能的情况下的一个折中选择。如果网络追求更快的速度，那么SE模块是一个更好的选择；如果需要更高的精度，那么在算力和内存允许的范围内保留更大尺度的特征图将能得到更精准的注意力。

4.5 与其他先进方法的比较

本文提出的CA-Net在GazeCapture数据集上与其他的眼睛凝视估计网络进行了对比，包括iTracker[12]，SAGE[13]，TAT[14]，AFF-Net[6]和InceptionResNet[15]。本文报告了这些模型在手机、平板上的物理平均误差，以及其用于凝视估计的眼睛图像的分辨率。结果如表2所列。CA-Net仅使用30×30像素(pixel)分辨率的眼睛图像进行凝视估计，是所有对比方法使用的眼睛图像分辨率最小的方法，在一定程度上验证了使用小分辨率眼睛图像进行凝视估计的可行性。此外，CA-Net在手机设备上实现了1.58 cm的精度，在平板设备上实现了2.13 cm的精度，均超越了目前最优的方法。

表2 CA-Net与其他先进方法的对比

Table 2 Comparison between CA-Net and other advanced methods

方法	手机凝视误差/cm	平板凝视误差/cm	眼睛图像分辨率/pixel
iTracker	1.86	2.81	224×224
SAGE	1.78	2.72	64×64
TAT	1.77	2.66	64×64
AFF-Net	1.62	2.30	112×112
InceptionResNet	1.59	—	128×128
CA-Net	**1.58**	**2.13**	**30×30**

4.6 类注意力编码

为了探究类注意力编码的作用，本文分别统计了测试集中所有参与者按尺度和眼睛维度进行分类时的平均类注意力编码，而由于按通道维度进行分类时的类注意力编码具有不可解释性，因此本文并未对其进行统计。尺度类注意力编码和眼睛类注意力编码分别由其对应模块中的最后一层线性层给出，其结果如表3和表4所列。

表3 平均尺度注意力编码

Table 3 Average scale attention coding

不同尺度特征所在层	平均注意力编码
第1层	0.37
第4层	0.39
第7层	0.61

表4 平均眼睛注意力编码

Table 4 Average eye attention coding

不同眼睛	平均注意力编码
左眼	0.48
右眼	0.70

如表3所列，本文分别统计了第1层、第4层和第7层卷积对应的平均尺度类注意力编码。根据实验结果可知，网络对不同尺度的特征分配的注意力并不相同，越深层的特征被分配了越高的注意力权重。这表明深层特征在网络进行凝视估计的过程中的贡献比浅层特征更大，但同时浅层特征也不并不是毫无作用，其仍对最终的凝视估计有所贡献。

由表4可以发现右眼被分配的注意力权重远比左眼高，这与人们通常以右眼作为主视眼和以左眼辅助进行注视的现象相符。眼睛类注意力模块可以根据不同的眼睛特征自适应地提取眼睛类注意力编码，融合眼睛类注意力编码可以让网络更合理地利用左眼和右眼的信息，进而得到更精准的凝视估计位置。

结束语 本文引入了一种基于RGB外观的凝视估计方法，命名为类注意力网络(CA-Net)。CA-Net通过按通道、尺度和眼睛逐类地提取和融合类注意力编码到每个特征中，

增加了与凝视相关特征的权重，进而显著提升了在 GazeCapture 数据集上的凝视估计精度。此外，本文通过改进压缩激励(SE)模块，提出了轻压缩激励模块，该模块在牺牲小部分算力的情况下获得了更高精度的通道注意力。在实验部分，本文通过消融实验详细对比验证了各个类注意力模块和设备种类信息的有效性，并且和其他先进工作进行了对比。最后讨论和分析了类注意力编码的含义。

通过实验结果和讨论，本文验证了 CA-Net 在凝视估计任务中的卓越性能，所提出的类注意力网络为凝视估计领域带来了新的思路和方法，通过多层次、多类别的注意力编码，使网络能够更好地理解和利用输入数据的特征，从而实现更准确的凝视点估计。未来的研究可以进一步探索和拓展这一方法，在更多复杂场景和实际应用中验证其通用性。

参 考 文 献

[1] GUO B Y，FANG W N. Fatigue detection method based on eye tracker[J]. Aerospace Medicine and Medical Engineering，2004(4)：256-260.

[2] YAN G L. The Application of Eye Movement Analysis in Advertising Psychology Research[J]. Psychological Dynamics，1999(4)：50-53.

[3] BORGESTIG M，SANDQVIST J，AHLSTEN G，et al. Gaze-based assistive technology in daily activities in children with severe physical impairments-an intervention study[J]. Pediatric Rehabilitation，2017，20(3)：129-141.

[4] CHENNAMMA H R，YUAN X H. A survey on eye-gaze tracking techniques[J]. Indian Journal of Computer Science and Engineering，2013，4(5)：388-393.

[5] MURTHY L R D，PRADIPTA B. Appearance-based gaze estimation using attention and difference mechanism[C]// Computer Vision and Pattern Recognition Workshops. IEEE Computer Society，2021：3137-3146.

[6] BAO Y，CHENG Y，LIU Y，et al. Adaptive feature fusion network for gaze tracking in mobile tablets[C]// International Conference on Pattern Recognition. The International Association for Pattern Recognition，2021：9936-9943.

[7] HU J，SHEN L，SUN G. Squeeze-and-excitation networks [C]// Computer Vision and Pattern Recognition. IEEE Computer Society，2018：7132-7141.

[8] GUESTRIN E D，EIZENMAN M. General theory of remote gaze estimation using the pupil center and corneal reflections [J]. IEEE Transactions on Biomedical Engineering，2006，53(6)：1124-1133.

[9] SUN L，LIU Z，SUN M T. Real time gaze estimation with a consumer depth camera[J]. Information Sciences，2015(320)：346-360.

[10] DEMENTHON D F，DAVIS L S. Model-based object pose in 25 lines of code[J]. International Journal of Computer Vision，1995，15(1/2)：123-141.

[11] WANG K，JI Q. Real time eye gaze tracking with 3d deformable eye-face model[C]// International Conference on Computer Vision. IEEE Computer Society，2017：1003-1011.

[12] KRAFKA K，KHOSLA A，KELLNHOFER P，et al. Eye tracking for everyone[C]// Computer Vision and Pattern Recognition. IEEE Computer Society，2016：2176-2184.

[13] HE J F，PHAM K，LALLIAPPAN N，et al. On-device few-shot personalization for real-time gaze estimation[C]// International Conference on Computer Vision Workshop. IEEE Computer Society，2019：1149-1158.

[14] GUO T C，LIU Y C，ZHANG H，et al. A generalized and robust method towards practical gaze estimation on smart phone[C]// International Conference on Computer Vision Workshop. IEEE Computer Society，2019：1131-1139.

[15] ATHAVALE R，MOTATI L S，KALAHASTY R. One eye is all you need：lightweight ensembles for gaze estimation with single encoders[J]. arXiv：2211. 11936，2022.

[16] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017：6000-6010.

[17] SHI W Z，CABALLERO J，HUSZAR F，et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network[C]// Computer Vision and Pattern Recognition. IEEE Computer Society，2016：1874-1883.

[18] JACOB D，CHANG M W，LEE K，et al. BERT：pre-training of deep bidirectional transformers for language understanding [C]// North American Chapter of the Association for Computational Linguistics：Human Language Technologies. 2019：4171-4186.

[19] PORAC C，COREN S. The dominant eye[J]. Psychological Bulletin，1976，83(5)：880-897.

[20] BAO J，LIU B，YU J. An individual-difference-aware model for cross-person gaze estimation[J]. IEEE Transactions on Image Processing，2022，31：3322-3333.

[21] KARTYNNIK Y，ABLAVATSKI A，GRISHCHENKO I，et al. Real-time facial surface geometry from monocular video on mobile gpus[J]. arXiv：1907. 06724，2019.

XU Jinlong，born in 1985，Ph.D，master's supervisor. His main research interests include high performance computing and parallel compilation.

HAN Lin，born in 1978，Ph.D，associate professor，is a senior member of CCF (No. 16416M). His main research interests include compiler optimization and high performance computing.

（责任编辑：何杨）