计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(10):66−74 [doi: 10.15888/j.cnki.csa.009649]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

双向融合 CNN 与 Transformer 的三维视线估计①

吕嘉琦, 王长元

(西安工业大学 计算机科学与工程学院, 西安 710021)
通信作者: 王长元, E-mail: cyw901@163.com

摘　要: 针对当前视线估计任务在无约束环境中易受影响因素干扰, 准确度不高的问题, 提出一种卷积与注意力双分支并行的特征交叉融合视线估计方法, 提升了特征融合的有效性和网络性能. 首先, 对 Mobile-Former 网络进行改进, 引入了线性注意力机制和部分卷积, 有效提高了特征提取能力并且降低了计算成本; 其次, 增加了基于 300W-LP 数据集预训练的 ResNet50 头部姿态特征估计网络分支来增强视线估计的准确度, 并使用 *Sigmoid* 函数作为门控单元来筛选有效特征; 最后, 将面部图像输入神经网络进行特征提取和融合, 输出三维视线估计方向. 在 MPIIFaceGaze 和 Gaze360 数据集上评估模型, 该方法的视线平均角度误差为 3.70°和 10.82°, 通过与其他主流三维视线估计方法比较, 验证了该网络模型能够比较准确的估计三维视线方向并降低计算复杂度.

关键词: 三维视线估计; 并行结构; 双向融合; 部分卷积; 线性注意力机制

引用格式: 吕嘉琦,王长元.双向融合 CNN 与 Transformer 的三维视线估计.计算机系统应用,2024,33(10):66–74. http://www.c-s-a.org.cn/1003-3254/9649.html

3D Gaze Estimation by Bidirectional Fusion of CNN and Transformer

LYU Jia-Qi, WANG Chang-Yuan

(School of Computer Science and Engineering, Xi'an Technological University, Xi'an 710021, China)

Abstract: To address the issue of low accuracy and susceptibility to interference from external factors in unconstrained environments, a convolution and attention double-branch parallel feature cross-fusion gaze estimation method is proposed to enhance feature fusion effectiveness and network performance. Firstly, the Mobile-Former network is enhanced by introducing a linear attention mechanism and partial convolution. This effectively improves the feature extraction capability while reducing computing costs. Additionally, a branch of the ResNet50 head pose feature estimation network, pre-trained on the 300W-LP dataset, is added to enhance gaze estimation accuracy. A *Sigmoid* function is used as a gating unit to screen effective features. Finally, facial images are inputted into the neural network for feature extraction and fusion, and the 3D gaze estimation direction is outputted. The model is evaluated on the MPIIFaceGaze and Gaze360 datasets, and the average angle error of the proposed method is 3.70° and 10.82°, respectively. The network model is verified to accurately estimate the 3D gaze direction and reduce computational complexity compared to other mainstream 3D gaze estimation methods.

Key words: 3D gaze estimation; parallel structure; bidirectional fusion; partial convolution; linear attention mechanism

三维视线估计是计算机视觉领域的一个关键任务, 它旨在通过对二维图像或视频中的人类视线的跟踪和估计推断出人物在三维空间中的注视方向. 该任务对于许多应用领域具有重要意义, 如增强现实、虚拟现

① 基金项目: 国家自然科学基金 (52072293)
收稿时间: 2024-03-18; 修改时间: 2024-04-16; 采用时间: 2024-05-14; csa 在线出版时间: 2024-08-21
CNKI 网络首发时间: 2024-08-22

实、自动驾驶等. 目前, 研究人员提出了各种各样的方法来解决三维视线估计的问题, 目前主要分为两类: 基于几何模型的方法和基于外观的方法[1]. 基于模型的方法通过预先建立的人眼模型和图像特征点, 推断三维视线方向, 这些方法需要专门的设备来捕获特定的眼睛信息. 因此, 基于外观的方法引起了人们的广泛关注. 基于外观的方法只需要一个摄像头来捕获图像, 以人眼或面部图像作为输入, 直接学习从面部外观到注视方向的映射函数, 并且表现出较好的视线估计结果. 然而, 由于图像中存在的多种挑战, 如遮挡、光照变化和噪声等, 三维视线估计仍然是一个具有挑战性的问题[2].

深度学习技术的进步, 推动了许多基于外观的深度神经网络模型的发展, 以解决三维视线估计任务. 2015 年, Zhang 等人[3]提出一个类似 LeNet 的浅层网络, 以单只眼睛的图像为输入, 并将头部角度向量加在全连接层的输出上, 这是第 1 次将 CNN 应用于三维视线估计, 其视线估计性能超越了大多数传统的基于外观的视线估计方法, 但检测到人脸后还需要再去检测眼部区域, 存在较多图像预处理步骤. Cheng 等人[4]受到注视估计中左右眼不对称现象的启发, 提出了一个非对称回归评估网络, 通过评估两只眼睛的表现来自适应调整以提高视线估计性能. 2016 年, Krafka 等人[5]提出了 iTracker 卷积神经网络用于 iPhone 和 iPad 用户的视线估计, 构建了用于视线估计的数据集 Gaze Capture. 2019 年 Chen 等人[6]提出空洞卷积网络 Dilated-Net, 使用空洞卷积在不降低空间分辨率的情况下提取高级特征来提高三维视线估计的准确性. 上述基于 CNN 的方法提高准确性不可避免地伴随着网络层数加深, 模型复杂度变大等问题, 并且其全局建模能力有限, 难以进一步提高视线估计性能.

ViT (vision Transformer)[7]采用了 Transformer 结构, 基于自注意力机制, 能够更好地捕捉图像中的长距离依赖关系, 使得在处理全局信息和复杂任务时表现优秀. 2022 年 Cheng 等人[8]提出了 GazeTR 网络, 首次使用 Transformer 结构模型用于视线估计, 首先利用 CNN 从局部特征图中提取特征然后使用 Transformer 编码器从特征图中估计视线方向. Transformer 的自注意力机制的计算和存储随着空间维度呈二次增长, 从而带来巨大的计算成本.

其后 Li 等人[9]提出将卷积结构与 Swin Transformer 采用串行结构相结合的混合视线估计模型 Res-Swin-GE, 但是串行结构对特征融合的程度非常有限, 且不能很好地保留局部空间特征.

因此, 本文基于 Mobile-Former[10]对模型进行改进, 改进的网络包括 3 条分支, 以人脸图像为输入的 CNN 与头部姿态特征分支, 随机初始化的可学习 token 输入 Transformer 分支. CNN 与 Transformer 分支对图像进行多次特征提取和融合, 最后再与头部分支的特征融合后回归出三维视线方向. 实验结果表明, 在 MPIIFace-Gaze 和 Gaze360 数据集上使用改进的 Mobile-Former 进行视线估计时, 与其他方法进行对比实验, 其精度表现最好, 并且在计算复杂度方面具有优势. 主要工作如下.

(1) 在 Mobile-Former 的基础上, 引入部分卷积 (partial convolution)[11]与线性注意力机制 AFT (attention free Transformer)[12], 分别替换 CNN 分支的深度卷积和 Transformer 分支的多头自注意力机制, 使得网络能够有效降低浮点运算次数并提高特征提取能力.

(2) 引入基于 ResNet50[13]的头部姿态特征提取分支作为第 3 分支, 在 300W-LP 数据集[14]上进行预训练, 该分支进一步提高了模型的表达能力, 并且在 CNN 和头部分支加入了基于 *Sigmoid* 函数的门控单元, 能够高效的筛选有效特征信息.

1 本文方法

在无约束的环境中, 基于外观的视线估计将会面对更多挑战, 例如头部姿态、个人差异和环境影响. 这些因素对面部外观有很大影响, 使得采集的图像信息包含复杂的影响因素. 为了很好地应对这些复杂条件, 本文基于改进的 Mobile-Former 网络进行视线估计. 本文的模型结构如图 1 所示, 将面部图像送入 CNN 分支和头部姿态分支, 随机初始化的 token 送入 Transformer 分支, 然后进行特征提取. CNN 和 Transformer 分支使用 Mobile-Former 的交叉融合模块进行多轮特征融合, 最终对 3 个分支的特征进行拼接, 并通过两个全连接层回归出三维视线向量, 将其转换为三维视线方向.

1.1 改进的 Mobile-Former 特征提取

Mobile-Former 是 MobileNet[15]和 Transformer 的并行结构, 两个分支使用中间的双向桥进行特征融合. 这种结构利用了 MobileNet 和 Transformer 的优点, 即 CNN 在提取局部特征的效率及 Transformer 在全局建

模方面的能力, 实现了局部和全局特征的双向融合. 在 CNN 分支, 以图像作为输入, 使用深度可分离卷积, 引入倒置瓶颈块结构来提取局部特征. Transformer 分支以可学习的 token 作为输入, 以多头自注意力机制进行全局信息建模. 此外, 这两个分支通过双向融合局部和全局特征的交叉融合模块进行通信.

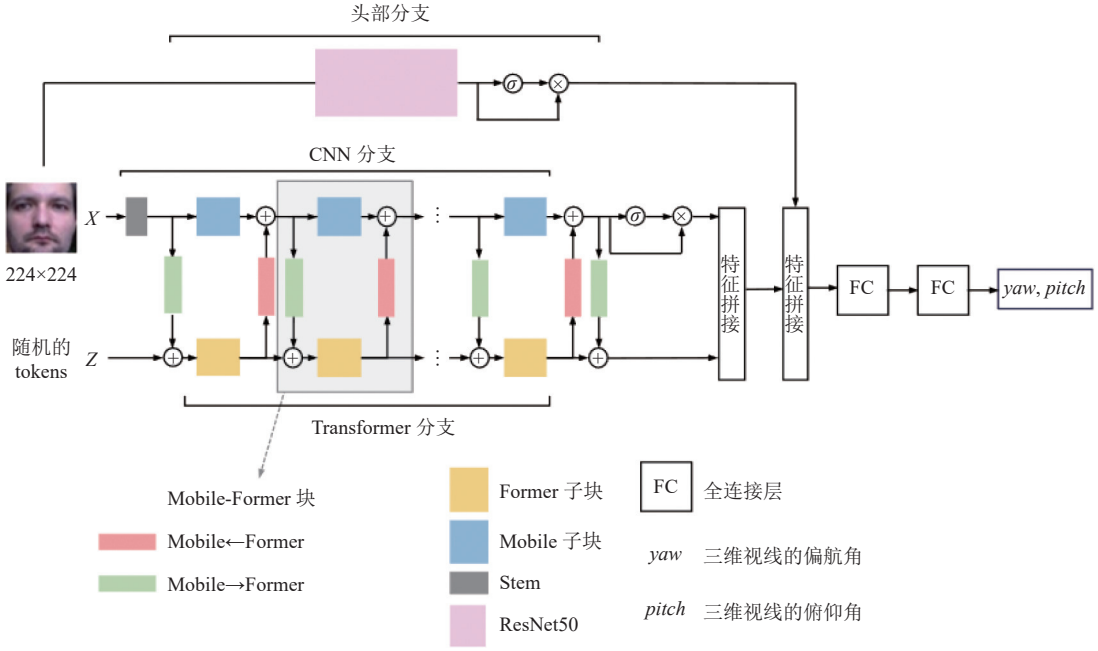


图 1 改进的 Mobile-Former 网络结构图

1.1.1 双向桥模块

双向桥是 Mobile-Former 中 CNN 与 Transformer 分支进行局部与全局特征双向融合的轻量交叉注意力模块, 其中有两个方向分别表示为 Mobile→Former 和 Former→Mobile, 如图 2 所示. Mobile→Former 用于将局部特征融合到全局特征中, 计算公式为:

$$A_{x\to z}=\bigcup_{i=1}^{h} Attn(\widetilde{z}_i W_Q^i, \widetilde{x}_i, \widetilde{x}_i) W_O \tag{1}$$

其中, $\widetilde{z}_i$表示第i个全局 token, W_Q^i是该头的查询投影矩阵, $\widetilde{x}_i$是局部特征图, W_O用于将多头的输出合并, $Attn(Q,K,V)$是标准的注意力函数.

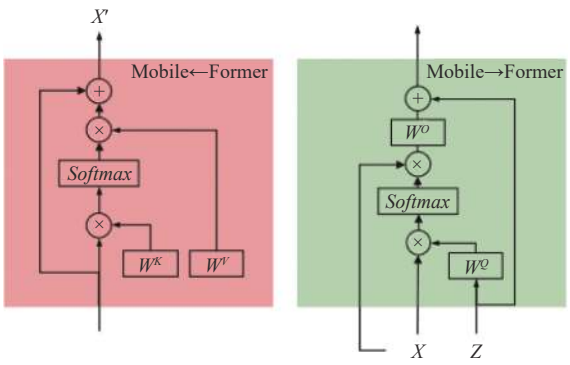


图 2 双向桥模块

从 Former→Mobile 的注意力可表示为:

$$A_{z\to x}=\bigcup_{i=1}^{h} Attn(\widetilde{x}_i, \widetilde{z}_i W_K^i, \widetilde{z}_i W_V^i) \tag{2}$$

其中, $\widetilde{z}_i$仍然是全局 token, W_K^i和W_V^i分别是键 (K) 和值 (V) 的投影矩阵.

由式 (1) 可以看出, 在 Mobile→Former 的方向上, 通过移除 Mobile 侧的键 (K) 和值 (V) 的投影矩阵, 以节省计算量, 这使得模型在保持较低 FLOPs 的同时, 能够有效地融合局部和全局信息, 提高模型性能.

1.1.2 Transformer 分支

该分支由多个 Former 子块组成, 这是一个标准的 Transformer 块, 包括多头注意力机制和前馈神经网络, 它的输出用于生成 Mobile 子块中的动态 ReLU[16]参数. 本文改进了 Mobile-Former 网络, 以更好地提取图像特征并降低模型的复杂度, 加入了线性注意力机制 AFT, 替换掉 Transformer 分支的标准多头自注意力机制, 改进后的 Former 块结构如图 3 所示.

这个分支以可学习的参数$Z\in R^{M\times d}$为输入, 其中 M 和 d 分别是参数的数量和维度. 这些参数是随机初始化的, 用来作为图像的全局先验知识. 整个分支以堆叠带有多头注意力 (multi-head self attention) 和前馈网

络 (feed forward network, FFN) 的标准 Transformer 块组成. Transformer 块中的关键是自注意力机制, 用于提取图像中的视线特征并自动关注图像中的不同部分, 自注意力机制的核心是查询权重矩阵 $W_Q \in R^{T\times d_k}$, 键权重矩阵 $W_K \in R^{T\times d_k}$ 和值权重矩阵 $W_V \in R^{T\times d_v}$ 的线性变换以及注意力权重的计算, 其中 T 是序列的长度, d_k 和 d_v 是每个特征的维度, 自注意力机制最终可表示为:

$$Attn(Q,K,V) = Softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{3}$$

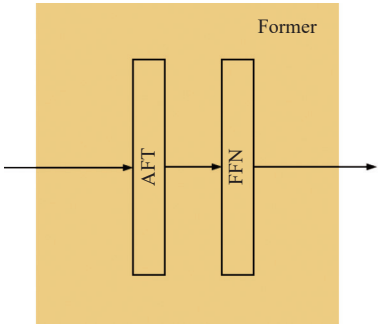


图 3 改进的 Former 块

多头注意力机制通过并行多个独立的注意力头将输入序列映射到多个查询、键和值空间, 并同时计算多个注意力头的权重, 然后将它们的输出拼接在一起, 从而能全面地捕捉序列中的信息关系.

为了降低模型的计算复杂度并提高三维视线估计精度, 本文改进了 Transformer 块. 原始网络中的自注意力机制时间和空间复杂度都很高, 首先需要初始化 3 个权重矩阵, 之后用输入序列乘以对应的权重矩阵, 接下来, 使用点积运算计算每个元素的查询和所有元素的键之间的相似度, 最后计算注意力权重并加权求和. 在输入序列长度为 T, token 长度为 d 的情况下, 多头注意力机制的时间复杂度为 $O(T^2d)$, 空间复杂度为 $O(T^2+Td)$. 本文引入了 AFT 替代了所有 Transformer 块中的自注意力模块.

AFT 是一种高效的注意力机制模块, 它消除了标准注意力机制的点积运算, 这使得 AFT 的操作具有线性的计算复杂度, 适用于大型的输入与模型尺寸, 它在语言模型建模, 图像分类等任务上取得了与标准注意力机制和其他变体注意力机制相当的性能, 同时提供更出色的效率, 这为注意力模型的设计思路开辟了新的设计空间, 对需要注意力机制的各个领域产生影响[12].

AFT 的主要计算思想为: 输入 X 经过 3 个线性变化得到 Q, K, V 这 3 个矩阵, 其维度为 $R^{T\times d}$, T 为序列长度, d 是每个特征的维度. AFT 引入了一个可训练的一对一位置偏执矩阵 $w \in R^{T\times T}$, 对于每个目标位置 t, 累加 $w_{t',t}$ 与 $K_{t'}$ 并使用 *Softmax* 归一化, 可得:

$$Weighted(K_{t'}) = \frac{\exp(K_{t'}+w_{t',t})}{\sum_{t'=1}^{T}\exp(K_{t'}+w_{t',t})} \tag{4}$$

使用 *Sigmoid* 激活函数对 Q_t 归一化, 然后点乘式 (4):

$$Attention_t = \sigma(Q_t)\odot\frac{\exp(K_{t'}+w_{t',t})}{\sum_{t'=1}^{T}\exp(K_{t'}+w_{t',t})} \tag{5}$$

最后根据权值对 V 进行加权运算:

$$Y_t = \sum_{t'=1}^{T}\sigma(Q_t)\odot\frac{\exp(K_{t'}+w_{t',t})}{\sum_{t'=1}^{T}\exp(K_{t'}+w_{t',t})}\odot V_{t'} \tag{6}$$

整理可得:

$$Y_t = \sigma(Q_t)\odot\frac{\sum_{t'=1}^{T}\exp(K_{t'}+w_{t',t})\odot V_{t'}}{\sum_{t'=1}^{T}\exp(K_{t'}+w_{t',t})} \tag{7}$$

总体而言, AFT 的核心算法没有使用矩阵乘法, 只使用了向量点乘和累加运算, 权重由键和可学习的位置偏差组成, 不需要计算和存储庞大的注意力矩阵, 同时还保持了查询和值之间的全局关系, 空间复杂度降低到了 $O(Td)$, 虽然时间复杂度仍是 $O(T^2d)$, 但计算量已有下降.

改进后的 Transformer 分支中, 拥有多个 AFT 注意力模块和前馈网络层 FFN, 对于输入 x, 经过一个 Transformer 块可表示为:

$$x' = \mathrm{FFN}(\mathrm{AFT}(x)) \tag{8}$$

其中, AFT(·)为线性注意力函数模块, FFN(·)为前馈神经网络映射函数, 输出 x' 流向交叉融合模块送往 CNN 分支. 将该分支最终输出特征展平为 1×192 的特征图.

1.1.3 CNN 分支

Mobile-Former 中以 MobileNet 作为 CNN 分支, 可拆分为堆叠多个 Mobile 子块组成, 在子块中, 主要由深度可分离卷积提取特征, 深度卷积的卷积核大小为 3×3, 对输入的特征图进行深度卷积, 然后对每个通

道进行逐点卷积, 并将 ReLU 替换为动态 ReLU, 如图 4 所示.

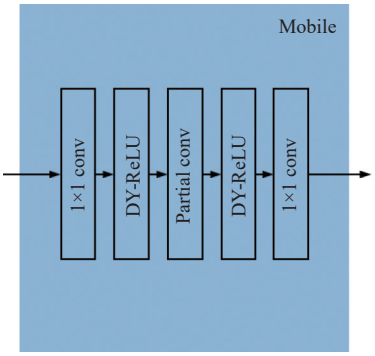


图 4 改进的 Mobile 块

为了提高模型性能并增强特征提取能力, 本文改进了 Mobile 子块, 加入部分卷积, 替换了原本的深度卷积. 深度卷积现在已经被广泛用作许多神经网络的关键模块, 对于输入$I \in R^{c\times h\times w}$, 深度卷积使用$c$个卷积核$w \in R^{k\times k}$来计算输出, 每一个卷积核只在一个通道上进行滑动. 对于计算复杂度, 主要以浮点运算 (FLOPs) 来衡量, 常规卷积的 FLOPs 为:

$$h\times w\times k^2\times c^2 \tag{9}$$

虽然深度卷积的 FLOPs 为:

$$h\times w\times k^2\times c \tag{10}$$

但其后通常跟着逐点卷积来提高通道数, 缓解通道降低带来的精度下降. 部分卷积是一种新型的卷积操作, 旨在通过减少计算冗余来提高神经网络的运行速度, 它只在输入通道一部分上使用常规卷积提取特征, 保持其余通道不变, 从而减少了计算冗余和内存访问, 可以考虑使用开头和结尾的几个连续通道c_p, 则部分卷积的 FLOPs 为:

$$h\times w\times k^2\times c_p^2 \tag{11}$$

根据式 (11), 本文模型设置为取一半的通道应用常规卷积, 则部分卷积的 FLOPs 仅为常规卷积的 1/4, 从而实现了更高的计算速度.

并且使用部分卷积替代深度卷积, 不用更换后面的逐点卷积, 这样的组合形式使得在特征图上有效的感受野像一个 T 型, 更专注于中心的位置[11], 这对于双分支并行的结构来说, 更加发挥 CNN 分支对于局部特征提取的特点, 如图 5 所示. 总的来说, 使用部分卷积只需要从一部分通道中提取特征, 后接逐点卷积, 便可以充分利用所有通道的信息并提高性能. 本文设置仅有一半的连续通道进行卷积操作. 该分支输出特征展平为 1×1152, 经过门控单元后与 Transformer 分支特征拼接, 再经过全连接层变为 128 通道.

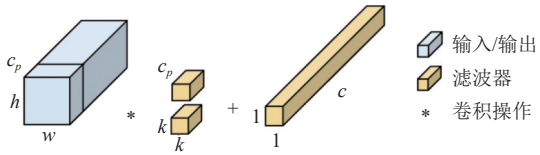


图 5 部分卷积与逐点卷积组合的 T 型结构

1.2 头部特征分支

在无约束的环境下, 头部姿态对视线方向有很大的影响. 人类在观察对象时通常会先转动头部朝向目标, 然后再调整眼睛的注视方向. 因此, 考虑头部姿态可以提供额外的上下文信息, 有助于更精确地预测用户的视线方向. 特别是在一些极端情况下, 如大幅度的头部转动或视线方向与头部方向不一致时, 头部姿态分支能够纠正原始的视线方向, 提高了模型的准确性和鲁棒性.

本文在改进的 Mobile-Former 基础上引入第 3 分支, 用来估计头部姿态特征以增强视线方向的准确性. 头部特征分支接收和其他分支一样的面部图像作为输入, 以标准的 ResNet50 作为神经网络, 在 300W-LP 数据集上进行预训练, 这个数据集包含大量的低分辨率人脸图像, 模拟了真实世界中移动设备和监控摄像等低分辨率的场景, 其中包括各种姿态、光照、面部表情等影响因素, 从而弥补了视线方向估计数据集人脸图像场景、环境因素较为单一的缺点.

Deng 等人[17]认为视线方向作为一个整体, 是由头部姿态和眼球运动共同决定的, 二者之间存在确定性的几何关系, 因此本文不再将头部分支最终输出头部姿态欧拉角 (俯仰角、偏航角、滚动角) 与视线特征直接拼接. 去掉 ResNet50 的最后的全连接输出层, 加入 1×1 卷积层将通道从 2 048 缩放为 64.

最后, 本文还加入了基于 *Sigmoid* 设计的门控单元, 如图 6 所示.

在门控单元中, 通过 *Sigmoid* 函数产生 0–1 之间的门控信号, 表示每个元素对应位置的重要程度, 然后将原始输入与门控信号相乘, 来选择性的保留或丢弃特定的输入特征, 使得网络可以自适应地提取和传递有效的信息, 从而提高模型的鲁棒性, 门控单元如式 (12):

$$x' = Sigmoid(x) \odot x \tag{12}$$

其中, ⊙表示逐元素乘积, *Sigmoid*(·)表示 *Sigmoid* 函数. 为本文将此门控单元加在 CNN 分支与头部特征分支的最后一层. 头部分支将 64 通道特征图通过门控单元后与 128 通道视线特征拼接为 192 通道, 最终通过全连接层输出视线方向, 即俯仰角与偏航角.

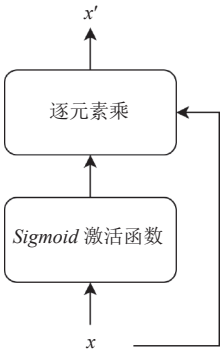


图 6 门控单元模块

2 实验与分析

2.1 数据集与评价指标

MPIIFaceGaze[18]和 Gaze360[19]三维视线估计方法中常用的数据集. 本文使用与其他最先进的视线估计方法相同的方法[3]对 MPIIFaceGaze 和 Gaze360 数据集进行归一化处理, 具体来说, 对虚拟相机进行平移和旋转, 去除头部滚动角, 保持虚拟相机与人脸中心之间的距离相同, 并在这两个数据集上训练和评估.

MPIIFaceGaze 数据集是有由 15 名受试者参与的数据集, 每个受试者有 3 000 张人脸图像. 评估采用留一交叉验证法, 即选取一个文件夹的图像作为测试集, 剩下的文件夹图像作为训练集, 然后对数据集中的每一个文件夹重复此过程, 最后对结果取平均值, 以评估模型的整体性能. 对于 Gaze360 数据集, 使用与 Gaze360 网络相同的方法[19]将整个数据集划分为训练集、测试集和验证集. Gaze360 数据集中有一些图像只显示受试者的背部, 所以这些图像不适合基于面部图像的视线估计检测, 使用 Phi-ai Lab 的方法, 删除没有人脸检测结果的图像, 以确保基于外观的视线估计方法的准确性和有效性. 图 7 显示了一些来自 MPIIFaceGaze 和 Gaze360 数据集中的示例图片.

通过改进的 Mobile-Former 网络最终可计算出俯仰角 (*pitch*) 和偏航角 (*yaw*), 之后该模型可以计算出代表注视方向的三维向量$\alpha = (x, y, z)$, 计算方法如下:

$$x = \cos(pitch)\cos(yaw) \tag{13}$$

$$y = \cos(pitch)\sin(yaw) \tag{14}$$

$$z = \sin(pitch) \tag{15}$$

在三维视线估计任务中最常用的评价指标是平均角度误差 (°), 即预测的注视方向和真实注视方向之间的角度. 平均角度误差 (°) 可由式 (11) 即计算:

$$E_{\text{angular}} = \arccos \frac{\alpha \cdot \beta}{|\alpha| \cdot |\beta|} \tag{16}$$

其中, ·表示两个向量的点积, α与β分别代表预测的注视方向和真实的注视方向, $|\alpha|$和$|\beta|$表示两个向量的大小, arccos 是反余弦函数.

(a) Gaze360 (b) MPIIFaceGaze

图 7 MPIIFaceGaze 和 Gaze360 中的一些图片

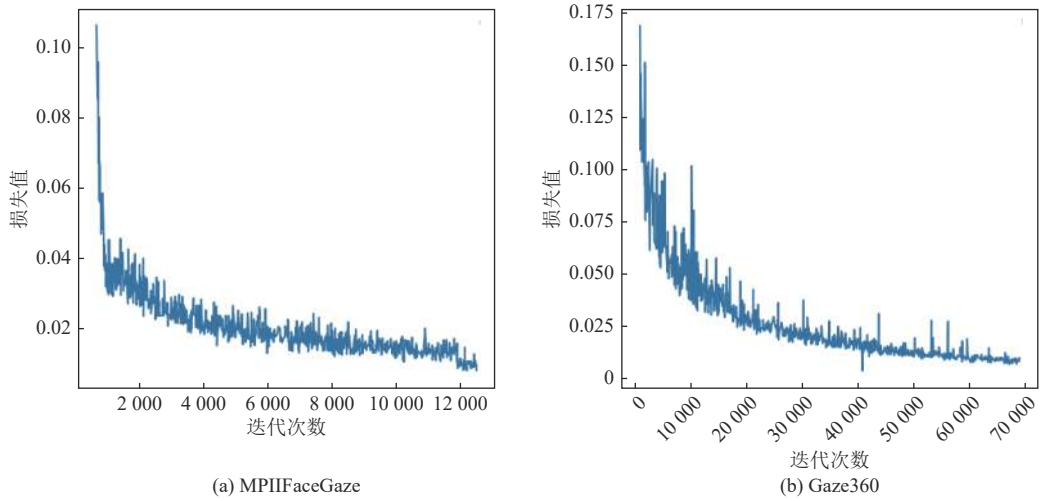
2.2 实验细节

本文的实验设备为配有 60 GB 内存, RTX 3090 GPU 的 Ubuntu 服务器, 使用 PyTorch 1.10 和 Python 3.8 开展实验. 对于 MPIIFaceGaze 数据集, 批量大小 (batch size) 设置为 64, 迭代周期 (epoch) 为 18, 初始学习率为 0.000 2, 权重衰减为 0.5, 从第 15 个 epoch 开始衰减, 使用线性学习率预热 2 个 epoch. 对于训练 Gaze360 数据集, 批量大小为 100, 初始学习率为 0.000 1, 衰减率为 0.97, 从第 2 个 epoch 开始衰减, 一共训练了 80 个 epoch. 在两个数据集上都使用 AdamW 优化器训练模型, $\beta_1 = 0.9$, $\beta_2 = 0.95$. 在这两个数据集上均使用 L1 Loss 作为损失函数, 计算公式如下:

$$L = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \widehat{y_i}\right| \tag{17}$$

图 8 展示了 MPIIFaceGaze 中受试者 p00 和 Gaze360 数据集的损失收敛曲线.

ResNet50 在 300W-LP 数据集上已有公共预训练权重, 训练了 25 个 epoch, 使用 Adam 优化器, 学习率为 0.000 01, $\beta_1 = 0.9$, $\beta_2 = 0.999$, 使用迁移学习作为整个网络的第 3 分支.



(a) MPIIFaceGaze (b) Gaze360

图 8 MPIIFaceGaze 与 Gaze360 损失收敛曲线

2.3 不同视线估计网络模型评估对比

为了评估三维视线估计模型的性能, 将本文模型与最先进的方法在平均角度误差方面进行对比实验. 在 MPIIFaceGaze 数据集上, 实验结果如表 1 所示, 本文的方法的平均角度误差为 3.70°, 比 L2CS-Net 高 0.22°. 表 1 中对比的模型中包含了纯 CNN、纯 Transformer 以及混合模型, 说明了本研究采用的方法在三维视线估计准确度方面优于其他方法.

此外, 本文还对 MPIIFaceGaze 数据集中的每个受试者提供了本文方法的平均角度误差, 并与 Dilated-Net 和 GazeTR 比较, 在 15 个受试者中, 与 Dilated-Net 对比, 本文的方法在 13 个受试者上实现了更高的三维视线精度, 与 GazeTR 相比有 9 个受试者精度更高, 结果如图 9 所示.

表 1 在 MPIIFaceGaze 数据集上的实验结果对比 (°)

模型	平均角度误差
MPIIGaze[20]	5.40
Dilated-Net[6]	4.80
CA-Net[21]	4.10
AGE-Net[22]	4.09
GazeTR[8]	4.00
L2CS-Net[23]	3.92
Res-Swin-Ge[9]	3.75
本文模型	3.70

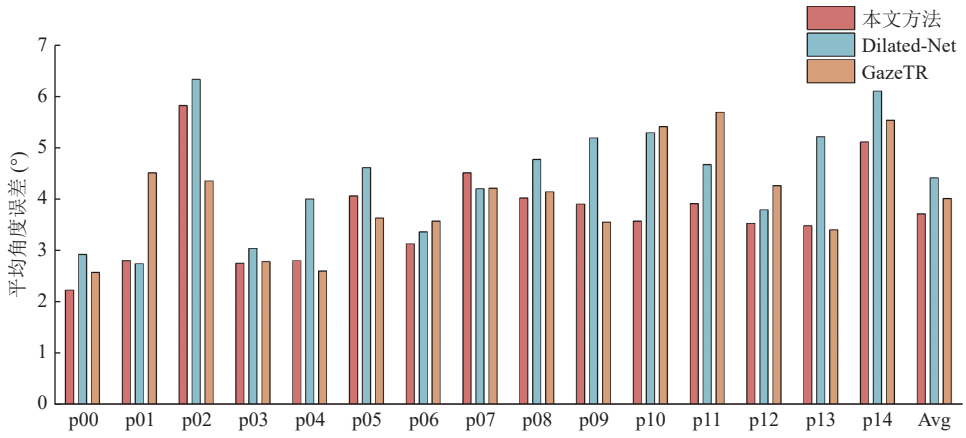


图 9 MPIIFaceGaze 数据集上的平均角度误差对比

表 2 显示了在 Gaze360 数据集的测试集上的对比实验结果. 在测试集上, 本文方法在前 180°的平均角度误差为 10.82°, 高于其他的视线估计方法.

为了验证改进的 Mobile-Former 模型在三维视线估计任务中的有效性和性能, 本文使用 PyTorch 的第三方库 THOP 统计了模型的参数量和浮点运算次数. 统计结果如表 3 所示.

参数量与计算复杂度代表着一个模型的空间和时

间复杂度, 而随着技术的不断进步, 开发更强大、显存更大的 GPU 变得相对容易. 这使得对于模型参数量和规模不再有过多的限制, 从而可以构建更加复杂的神经网络模型, 但这也导致更大的时间成本和能源消耗, 因此以空间换时间, 降低计算复杂度变得尤为重要.

表 2 在 Gaze360 测试集上的实验结果对比 (°)

模型	平均角度误差
Full-Face[18]	14.99
Dilated-Net	13.73
RT-Gene[24]	12.26
Gaze360	11.40
Bot2L-Net[25]	11.53
本文模型	10.82

表 3 不同模型性能对比

模型	平均角度误差 (°)	参数量 (M)	FLOPs (G)
Dilated-Net	4.40	3.920	3.153
GazeTR	4.00	11.394	1.834
本文模型	3.70	21.201	1.505

所以, 为了验证本文对模型在平均角度误差和浮点运算次数方面的有效性, 将本文方法与 Dilated-Net 和 GazeTR 进行对比, 实验结果如图 10 所示, 其中的离散点越靠近左下角, 浮点运算次数越少, 视线估计精度越高, 本文方法在这两个方面显著高于 Dilated-Net 与 GazeTR. 本文的改进虽然参数量稍有增加, 但在 MPIIFaceGaze 数据集上本文方法只训练了 18 个 epoch, 而 Dilated-Net 和 GazeTR 分别训练了 100 和 80 个 epoch, 充分说明了本文的改进在提升性能的同时有较强的拟合能力.

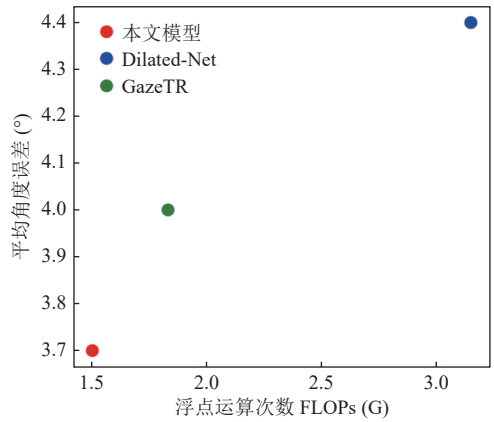


图 10 角度误差与计算复杂度的对比

2.4 头部姿态特征分支有效性验证

为验证加入头部姿态特征提取分支的有效性, 本文去除该分支, 仅保留两个视线特征分支, 在 MPIIFaceGaze 数据集上进行实验, 实验结果如表 4 所示. 由表 4 可以看出, 去掉头部分支后, 平均角度误差为 4.23°, 较去除头部分支前的模型有显著增加, 证明该分支学习到了 300W-LP 数据集中的极端头部姿态和环境因素影响, 对视线估计精度产生了一定的积极作用.

表 4 去除头部分支的实验结果比较 (°)

模型	平均角度误差
改进Mobile-Former	3.70
改进Mobile-Former (no head)	4.23

3 结束语

随着视线估计领域在生活中的大规模应用, 如何设计一个高精度且高性能的视线估计模型至关重要. 本文在 Mobile-Former 的 CNN 与 Transformer 并行交叉融合的启发下, 基于该模型进行改进, 加入部分卷积与线性注意力机制来减小计算复杂度, 设计并引入了头部姿态特征分支与门控单元, 增强了特征提取能力, 利用网络架构充分融合特征. 实验结果证明, 在 MPIIFaceGaze 与 Gaze360 数据集上, 通过与其他视线估计方法对比, 本文方法在两个数据集上以最低的平均角度误差实现了最高的注视精度, 并且保持了最低的浮点运算次数.

参考文献

1 苟超, 卓莹, 王康, 等. 眼动跟踪研究进展与展望. 自动化学报, 2022, 48(5): 1173–1192. [doi: 10.16383/j.aas.c210514]

2 Cheng YH, Bao YW, Lu F. PureGaze: Purifying gaze feature for generalizable gaze estimation. Proceedings of the 36th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2022. 436–443. [doi: 10.1609/aaai.v36i1.19921]

3 Zhang XC, Sugano Y, Fritz M, *et al*. Appearance-based gaze estimation in the wild. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 4511–4520. [doi: 10.1109/cvpr.2015.7299081]

4 Cheng YH, Lu F, Zhang XC. Appearance-based gaze estimation via evaluation-guided asymmetric regression. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 100–115. [doi: 10.1007/978-3-030-01264-9_7]

5 Krafka K, Khosla A, Kellnhofer P, *et al*. Eye tracking for everyone. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE,

2016. 2176–2184. [doi: 10.1109/cvpr.2016.239]
6 Chen ZK, Shi BE. Appearance-based gaze estimation using dilated-convolutions. Proceedings of the 14th Asian Conference on Computer Vision. Perth: Springer, 2019. 309–324. [doi: 10.1007/978-3-030-20876-9_20]
7 Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16x16 words: Transformers for image recognition at scale. Proceedings of the 9th International Conference on Learning Representations. OpenReview.net, 2021.
8 Cheng YH, Lu F. Gaze estimation using Transformer. Proceedings of the 26th International Conference on Pattern Recognition. Montreal: IEEE, 2022. 3341–3347. [doi: 10.1109/icpr56361.2022.9956687]
9 Li YJ, Chen JH, Ma JX, *et al*. Gaze estimation based on convolutional structure and sliding window-based attention mechanism. Sensors, 2023, 23(13): 6226. [doi: 10.3390/s23136226]
10 Chen YP, Dai XY, Chen DD, *et al*. Mobile-Former: Bridging MobileNet and Transformer. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 5270–5279. [doi: 10.1109/cvpr52688.2022.00520]
11 Chen JR, Kao SH, He H, *et al*. Run, don't walk: Chasing higher FLOPs for faster neural networks. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 12021–12031. [doi: 10.1109/cvpr52729.2023.01157]
12 Zhai SF, Talbott W, Srivastava N, *et al*. An attention free Transformer. Proceedings of the 2021 International Conference on Learning Representations. 2021. 1–15.
13 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/cvpr.2016.90]
14 Zhu XY, Lei Z, Liu XM, *et al*. Face alignment across large poses: A 3D solution. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 146–155. [doi: 10.1109/cvpr.2016.23]
15 Sandler M, Howard A, Zhu ML, *et al*. MobileNetV2: Inverted residuals and linear bottlenecks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 4510–4520. [doi: 10.1109/cvpr.2018.00474]
16 Chen YP, Dai XY, Liu MC, *et al*. Dynamic ReLU. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 351–367.
17 Deng HP, Zhu WJ. Monocular free-head 3D gaze tracking with deep learning and geometry constraints. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 3143–3152. [doi: 10.1109/iccv.2017.341]
18 Zhang XC, Sugano Y, Fritz M, *et al*. It's written all over your face: Full-face appearance-based gaze estimation. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu: IEEE, 2017. 51–60. [doi: 10.1109/cvprw.2017.284]
19 Kellnhofer P, Recasens A, Stent S, *et al*. Gaze360: Physically unconstrained gaze estimation in the wild. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 6912–6921. [doi: 10.1109/iccv.2019.00701]
20 Zhang XC, Sugano Y, Fritz M, *et al*. MPIIGaze: Real-world dataset and deep appearance-based gaze estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(1): 162–175. [doi: 10.1109/tpami.2017.2778103]
21 Cheng YH, Huang SY, Wang F, *et al*. A coarse-to-fine adaptive network for appearance-based gaze estimation. Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York: AAAI Press, 2020. 10623–10630. [doi: 10.1609/aaai.v34i07.6636]
22 Murthy LRD, Biswas P. Appearance-based gaze estimation using attention and difference mechanism. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 3143–3152. [doi: 10.1109/cvprw53098.2021.00351]
23 Abdelrahman A A, Hempel T, Khalifa A, *et al*. L2CS-Net: Fine-grained gaze estimation in unconstrained environments. Proceedings of the 8th International Conference on Frontiers of Signal Processing. Corfu: IEEE, 2023. 98–102. [doi: 10.1109/icfsp59764.2023.10372944]
24 Fischer T, Chang HJ, Demiris Y. RT-GENE: Real-time eye gaze estimation in natural environments. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 334–352. [doi: 10.1007/978-3-030-01249-6_21]
25 Wang XH, Zhou J, Wang L, *et al*. BoT2L-Net: Appearance-based gaze estimation using bottleneck Transformer block and two identical losses in unconstrained environments. Electronics, 2023, 12(7): 1704. [doi: 10.3390/electronics12071704]

(校对责编: 孙君艳)