

**虚拟现实技术基础**

**大作业**

**（2024—2025学年第一学期）**

**学 院 计算机科学与技术**

**班 级 16班**

**学 号 21210621**

**姓 名 马周原**

**任课教师 张振花**

**撰写日期 2024年11月30日**

深度学习视线估计算法：实现及剖析

摘要

视线估计算法是计算机视觉领域的重要研究方向，在人机交互、驾驶辅助和虚拟现实等应用中具有广泛的价值。本文系统地分析并复现了几种主流的深度学习视线估计算法，包括LeNet、ResNet-preact和Gaze-Net架构。在研究中，采用了该领域经典常用的 MPIIGaze数据集，并对数据进行了预处理，以提升模型训练和评估的性能。实验结果表明，深度网络架构在degree-error指标上优于浅层网络，尤其是在跨用户和跨设备的场景下展现出更强的泛化能力。

实验代码开源于：<https://github.com/fresh-ma/VR-GazeEstimation>

**关键词：**视线估计、深度学习、LeNet、ResNet、Gaze-Net

目录

**1 介绍**

* 1. 问题定义
  2. 主要应用

1. **研究现状**
   1. 相关工作
   2. 发展趋势
2. **算法**
   1. 基线模型
   2. ResNet-preact
   3. Gaze-Net
3. **实验**
   1. 数据集及预处理
      1. MPIIGaze
      2. 预处理
   2. 实验设定和结果
   3. 样例
4. **结论**

1 介绍

**1.1 问题定义**

视线估计（Gaze Estimation）是计算机视觉领域中的重要研究课题，其目标是通过分析眼睛或人脸图像，估算人的视线方向或注视点位置。得益于视线估计技术在众多应用场景中的中的重要性，如基于视线的人机交互[1]、视觉注意力分析[2, 3]、驾驶辅助以及虚拟现实等，该技术目前已经非常成熟[10]。

随着深度学习技术（Deep Learning）的飞速发展，基于学习的方法（learning-based methods）逐渐成为视线估计研究的主流。这些方法通过利用大规模的真实和合成数据集[4, 5, 6, 7]，实现了与个体无关的视线估计。然而，视线估计最终的归宿应该是高度无约束的，在这种场景下，视线估计仍面临诸多挑战，包括如何在日常环境中进行三维视线估计，以及如何在不依赖用户面部外观、环境几何特性和摄像头属性的条件下实现高精度估计。

尤其是在便携设备[8]和公共显示屏[9]等场景中，单目RGB摄像头的广泛使用为无约束视线估计提供了巨大的潜力，但也带来了噪声干扰、光照变化和多样化用户特性的难题。因此，研究针对单目RGB摄像头的无约束三维视线估计技术，不仅具有理论意义，还能够推动相关实际应用的发展。

本文主要聚焦于**以眼睛图像或人脸图像为处理对象，估算人眼视线方向**这一子任务。其形式化定义如下：

给定一幅输入图像 其中分别表示图像的高度、宽度和通道数。目标是通过一个映射函数：

从输入图像中预测视线方向向量，满足（单位向量）。

为了实现这一目标，优化过程通常基于以下损失函数进行建模：

其中，是真实视线方向， 是模型预测的视线方向，表示预测与真实视线方向之间的余弦相似度。

优化目标是通过最小化 来最大化预测视线与真实视线方向的相似度，从而提高估计精度。

**1.2 主要应用**

**基于视线的人机交互**（Gaze-based Human-Computer Interaction, Gaze-HCI）是视线估计技术最常见的应用之一。通过追踪用户的视线，系统可以捕捉用户的注意力、意图和情感状态，实现自然的交互方式。例如，用户可以通过眼动来控制计算机界面，选择目标或进行其他操作，从而替代传统的鼠标和键盘输入。这一技术对于老年人、残障人士等群体尤其重要，为他们提供了更便捷的交互方式。

**视觉注意力分析**（Visual Attention Analysis）涉及到理解人眼的注视模式，广泛应用于广告、教育、心理学研究和用户体验设计等领域。通过分析人眼的注视点和注视时间，可以有效评估人们在观看图像或视频时的兴趣焦点，从而优化广告内容、界面设计和学习材料的呈现方式。例如，广告商可以根据消费者的视线数据调整广告的布局，以提高广告效果和观众的参与度。

在**自动驾驶**和**驾驶辅助系统**中，视线估计技术被用来监测驾驶员的注意力状态，尤其是在疲劳驾驶检测中具有重要作用。通过实时分析驾驶员的眼睛和面部特征，系统可以判断驾驶员是否分心或疲劳，并在必要时发出警告，以提高行车安全性。这一技术的引入不仅能有效减少交通事故的发生，还能增强驾驶员的驾驶体验。

在**虚拟现实**（VR）中，视线估计技术能够增强用户的沉浸感和交互体验。通过精确追踪用户的视线，VR系统可以根据用户的注意方向动态调整显示内容，优化场景渲染，提高图像的质量和性能。此外，视线估计还可以用于VR中的情感识别与用户行为分析，从而为个性化的虚拟环境设置和互动设计提供数据支持。随着VR技术的快速发展，视线估计在这一领域的应用也日益广泛。

2 研究现状

**1.1 相关工作**

近年来，随着深度学习在众多领域获得显著成功，越来越多工作者把深度学习相关技术应用到了视线追踪任务上，取得了突破性的进展。

DL-based方法通常需要在大量眼睛图像上训练，使得模型能够学习眼睛外观和视线方向之间的复杂关系。这种方法通常对光照变化和部分遮挡具有较强的鲁棒性，但其性能依赖于训练数据的规模和质量，且需要较高的计算资源来进行模型推理。

最常见的步骤是使用头部或眼部图像来估计视线向量，特别是在没有三维交互的笔记本电脑应用中。Wong 等[14]提出了一种 ResNet 网络，通过输入一张经过色彩归一化的面部图像以及由偏航、滚转和俯仰角给出的头部姿态来推断视线向量。在驾驶场景中，定位并不总是准确，因此，先前的研究定义了必须估计的相关视线区域。Naqvi 等[15]创建了自己的驾驶数据集，其中包含17个不同的视线区域，并使用一个具有大量卷积的 CNN 来估计这些区域。另一方面，Illahi 等[16]提出通过仅使用归一化的 X、Y 屏幕坐标以及视线速度、头部旋转速度、视线加速度和头部旋转加速度来实时预测视线目标。这些变量作为输入给递归神经网络（RNN）。Katrychuk 等[11]比较了多层感知器（MLP）与浅层 CNN，使用从低功耗到高功耗的多种配置。这些设置旨在优化在如 Raspberry Pi 3 等低功耗系统中的实时跟踪。结果表明，浅层 CNN 在高功耗配置下获得了更好的结果（偏差为 0.55°），如预期的那样；然而，其他配置可能更适合加快训练和实时跟踪的速度。

另一些研究旨在从图像中估计瞳孔位置，这依赖于先前的校准，帮助将检测到的位置转换为视线向量[17]。Ou 等[17]使用 YOLOv3 网络，通过他们自己的近场可见光图像数据集预测瞳孔的中心。另一种方法是对图像进行语义分割，而不仅仅是检测瞳孔位置。Chaudhary 等[18]提出了一种类似 U-Net 的网络，与 SegNet 相比，能够对 OpenEDS 数据集进行语义分割。同样，Kothari 等[19]使用多个公开可用的数据集训练了一个编码器-解码器网络 DenseElNet，得出多个数据集的使用有效帮助获得更好的结果的结论。Chugh[20]描述了一种简单的 CNN，通过特征上采样和下采样，从单个眼睛的红外图像中估计瞳孔位置，获得了与预期输出的平均距离为 1.2 像素。Lu 等[21]训练了一个具有三层的简单 CNN，旨在去折射眼部图像并用五个参数（中心、轴和倾斜角）形成瞳孔的椭圆。他们在三维瞳孔的估计中达到了至少 2 毫米的误差。

另一个不太常见的设置是使用网络摄像头。在这种情况下，期望的输出通常是用户正在注视的像素。Gudi 等[22]将他们的方法论分为两个步骤：

（1）从输入图像中估计视线向量；

（2）将视线向量转换为屏幕上的视线位置。

第一步使用预训练的 VGG16 网络来解决，而第二步则采用三种方式进行解决：

（1）估计将一个空间转换为另一个空间的系数；

（2）使用机器学习；

（3）一种混合方法，在该方法中，机器学习被用来估计系数。

后者的表现优于其他两种方法，但仍然存在超过 4 厘米的误差。de Lope 和 Gran[12]检查了多个预训练网络在他们从笔记本电脑摄像头提取的数据集上的表现，发现 DenseNet 的准确率最高（在多用户数据集中为 91.30%）。与其让深度学习对普通公众变得复杂，不如存在能够优化网络及其超参数的框架[13]。通过这种方法，使用浅层 CNN 在 Columbia Gaze 数据集上获得了 85% 的准确率。

**1.2 发展趋势**

跨设备适应性：传统的视线估计方法通常依赖于特定硬件和数据集，因此其泛化能力较弱。近年来，研究者们越来越注重开发可以在不同设备（如手机、笔记本电脑、虚拟现实设备等）上广泛应用的视线估计模型。跨设备适应性已成为一个重要的研究方向，尤其是在考虑到智能设备在硬件性能和摄像头配置方面的差异时。为了实现更广泛的应用，研究者们开始使用迁移学习、领域自适应等技术，使得模型能够在多种设备上进行训练和推理，降低对特定硬件的依赖。

多模态融合：为了提高视线估计的准确性和鲁棒性，越来越多的研究开始探索将多种传感器数据进行融合。例如，结合眼睛图像、头部姿态、脸部特征以及环境信息（如光照、背景）等多模态数据进行视线估计。多模态融合不仅能够提高对复杂场景的适应能力，还能增强模型对遮挡、光照变化等因素的鲁棒性。通过多视角的输入，模型可以获得更全面的信息，从而提升对个体差异的适应能力。

实时和低延迟系统：视线估计技术的实时性对于许多应用至关重要，尤其是在虚拟现实（VR）、增强现实（AR）以及驾驶辅助系统等场景中。为了实现低延迟和高效的视线估计，研究者们不断优化模型结构，设计高效的推理算法。例如，轻量级神经网络（如MobileNet、ShuffleNet）和硬件加速（如GPU、FPGA）已被广泛应用于提高系统的实时性。通过这种方式，视线估计不仅能够满足对精度的需求，还能在实际应用中提供流畅的用户体验。

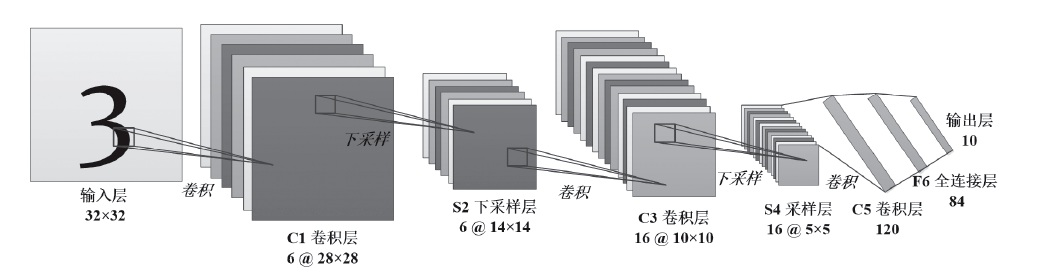
个性化视线估计：尽管目前的视线估计方法在大多数标准数据集上取得了较好的结果，但在个体差异较大的实际场景中，模型的泛化能力依然存在挑战。个性化视线估计成为另一个研究热点。通过对每个用户进行单独的训练或微调，研究者们力图提高系统在不同个体之间的适应能力。例如，通过利用少量的用户数据进行迁移学习或增量学习，系统可以根据用户的面部特征和眼睛形态动态调整模型，从而实现更加精确的视线追踪。

多任务学习：随着深度学习的多任务学习（MTL）方法的兴起，越来越多的研究开始将视线估计与其他相关任务（如面部表情识别、情感分析、头部姿势估计等）结合进行联合优化。这种方法通过共享不同任务间的特征表示，不仅可以提高视线估计的准确性，还能加速模型的训练过程，减少计算资源的消耗。多任务学习的引入，使得视线估计技术能够在多个任务之间找到有效的平衡，提高系统的综合性能。

3 算法

**1.1 LeNet**

LeNet 是由 Yann LeCun 等人在 1998 年提出的一种卷积神经网络（Convolutional Neural Network, CNN）结构，用于手写数字识别任务，是深度学习领域的开创性工作之一。LeNet 的结构设计较为简单，包含若干个卷积层、池化层和全连接层，其基本架构如下：



输入层：接收 32×32 的灰度图像。

卷积层：通过卷积操作提取图像的局部特征，使用小型滤波器进行空间特征提取。

池化层：采用平均池化（Average Pooling）对特征进行降维，从而减小计算量并增强模型的平移不变性。

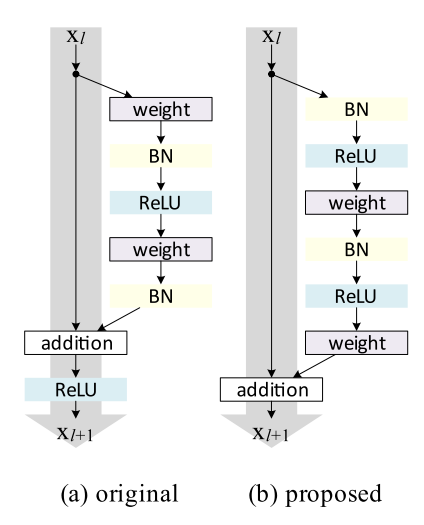
全连接层：将前面的特征映射到分类任务的目标空间。

输出层：使用 Softmax 激活函数进行多分类。

LeNet 的创新在于通过局部感受野和权值共享的方式显著降低了网络参数量，并增强了特征提取能力。虽然 LeNet 的结构相对简单，但为后续深度卷积神经网络的发展奠定了重要基础，在图像识别、语音处理等多个领域具有广泛的启发意义。

**1.2 ResNet-preact**

ResNet-preact 是 ResNet（Residual Network）的变体，由何恺明等人在 ResNet 的基础上提出，改进了残差块（Residual Block）的设计。与原始 ResNet 不同，ResNet-preact 在每个残差块中将批量归一化（Batch Normalization, BN）和激活函数（ReLU）移到了卷积层的前面，因此被称为“Pre-Activation ResNet”。



特点：

* 预激活设计：将 BN 和 ReLU 放在卷积操作之前，使梯度在反向传播过程中能够更顺畅地流动，有助于缓解梯度消失问题。
* 改进的梯度传播：预激活结构优化了梯度的分布，更适合深层网络。
* 简化的结构初始化：避免了由于激活函数位置导致的参数初始化问题，使得训练更高效。

ResNet-preact 的基本单元是预激活残差块，具体包括：

* 输入数据先经过 BN 和 ReLU，然后进行卷积操作。
* 添加跳跃连接（Skip Connection），将输入与卷积输出相加，从而实现残差学习。

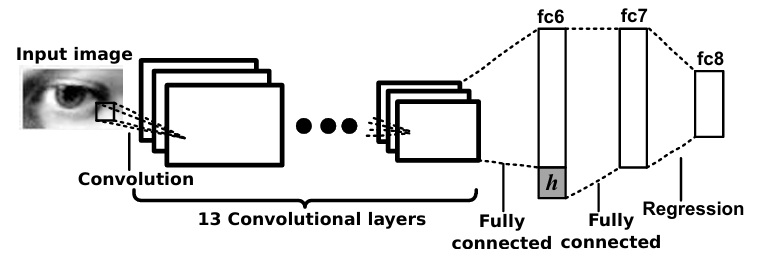
ResNet-preact 的提出进一步验证了残差学习在深度网络中的重要性，其改进结构在训练效率和模型性能上都优于原始 ResNet，特别适用于超过 100 层的深度网络。

**1.3 Gaze-Net**

Gaze-Net[10]本质上是一个基于预训练VGG-16的架构，以单张眼睛图像和头部姿态作为输入，注视向量作为输出。VGG（Visual Geometry Group）是由牛津大学的视觉几何组提出的一种卷积神经网络架构，VGG 网络在 2014 年的 ImageNet 挑战赛中表现出色，并被广泛应用于图像分类、目标检测等任务。VGG 的设计思想是通过增加网络的深度，采用非常简洁的结构来增强模型的表现力。其核心特点是使用小尺寸的卷积滤波器（3×3），并且堆叠多个卷积层来提取图像的层次特征。

特点：

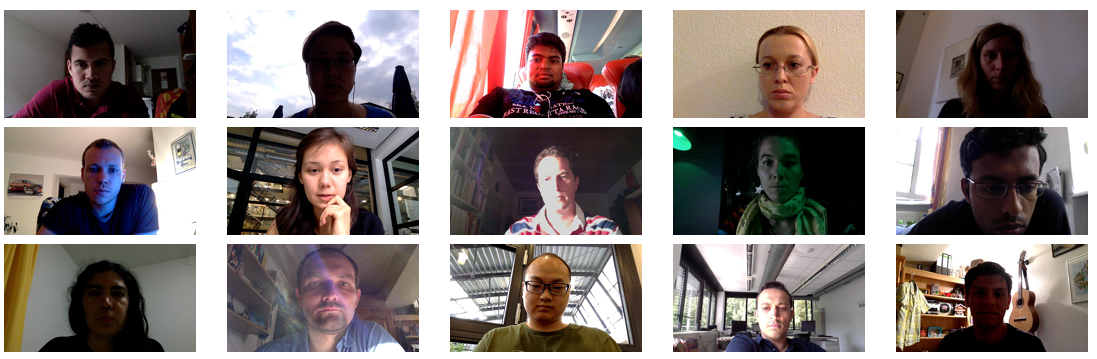
* 小尺寸卷积核：VGG 网络的主要创新之一是使用 3×3 的卷积核，这些小卷积核能够捕捉更精细的特征，并通过层叠的方式增强感受野。这与传统的大卷积核（如 5×5 或 7×7）相比，具有较少的参数量，从而减少了过拟合的风险。
* 深度堆叠结构：VGG 网络通过增加卷积层的深度来提高网络的表达能力。每个卷积块都包含多个 3×3 卷积层，最终输出一个高维的特征图。
* 统一结构：VGG 网络采用相同大小的卷积核和池化层，整个网络架构结构非常简洁且统一。网络的宽度（卷积层的通道数）逐渐增加，而深度（卷积层的数量）也逐渐加深。

****

4 实验

**1.1 数据集及预处理**

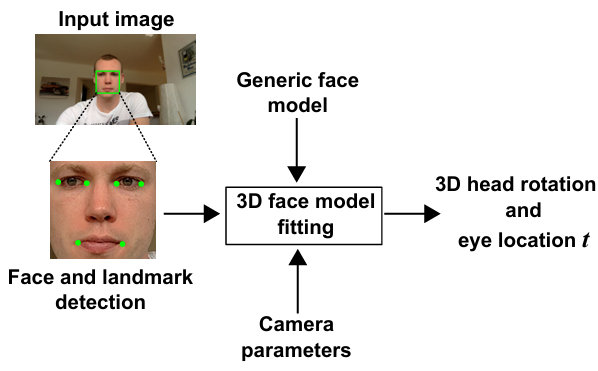
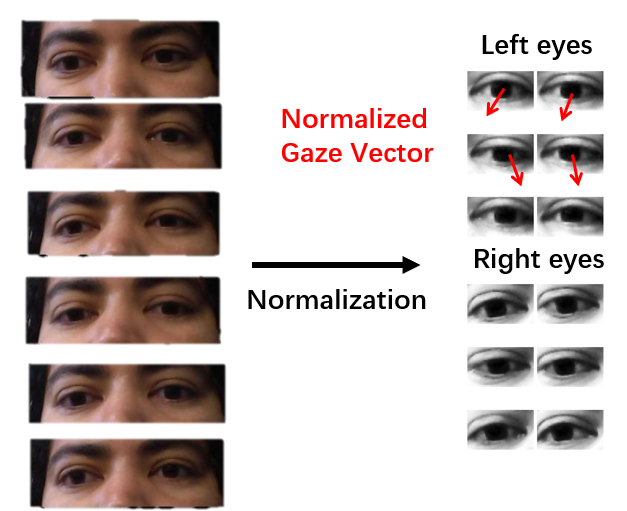
**MPIIGaze[10]**是一个用于自然环境下基于外观的视线估计的数据集。它包含从 15 名参与者在超过三个月的日常自然使用笔记本电脑期间收集的 213,659 张图像。数据集作者手工标注了37,667 张人脸6点landmark和瞳孔数据。如图1所示，该数据集在外观和光照方面具有很大的变化性。

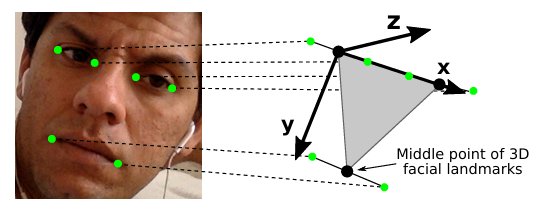
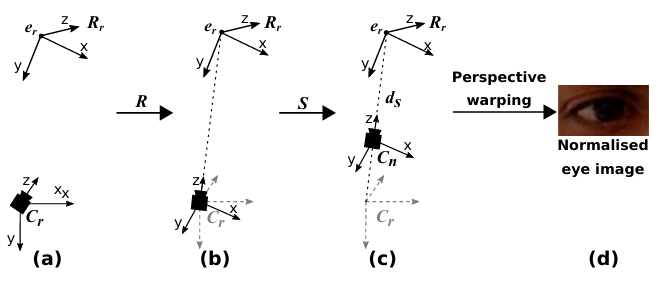
****

**图1：MPIIGaze数据集**

参考数据集作者原论文[10]的数据预处理方法，首先使用基于HOG的检测器对图像中的人脸进行检测，并假设每张图像中仅包含一个人脸。当检测到多个候选框时，保留面积最大的框；若未检测到人脸，则丢弃该图像（约占5%的数据）。随后，利用连续条件随机场（CCNF）模型检测面部关键点。

在三维头部姿态估计中，原作者采用了一个通用的平均面部形状模型F，其由所有参与者面部形状的平均值构建。该模型包含六个三维面部特征点（如眼角和嘴角），并通过EPnP算法估计初始姿态，再利用Levenberg-Marquardt方法进一步优化以拟合检测到的二维关键点。头部坐标系依据眼睛和嘴巴中点形成的三角形定义，确保姿态估计具有一致性和实用性。

**** 

**** ****

在我们的实验中，我们对于每个参与者随机选择了3000个样本（1500 个左眼样本和 1500 个右眼样本），共15\*3000=45000张图片。

**1.2 实验结果**

**实验环境：**

* **PyTorch** 2.3.0
* **Python** 3.12(ubuntu22.04)
* **Cuda**  12.1
* **GPU** RTX 3090(显存24GB) \* 1
* **CPU** 14 vCPU Intel(R) Xeon(R) Platinum 8362 CPU @ 2.80GHz
* **内存** 45GB

**实验设定细节：**

为了模拟真实应用环境种的测试，本实验对3个模型都进行了Cross-Person Evaluation，每次训练都将一个人的数据作为测试集、其余数据作为训练集。

指标选择Angle-Error，即计算ground-truth和预测向量之间的夹角。本指标越小代表预测越准确。

由于GazeNet内部的VGG-16训练耗时较长（在该环境下大约1分钟/epoch），我们并没有按照原论文训练60个epoch（耗时约15小时），而是在epoch=15时进行early stop（耗时约3.5小时）。**这可能导致从指标上看GazeNet的效果略逊于ResNet-preact-8。但理论上完全训练的GazeNet效果可以达到平均5.5左右[10]，只是我们没有训练至最优。**

在batch-size选择上，LeNet和ResNet选择较小的32，GazeNet选用256（参考原论文）。

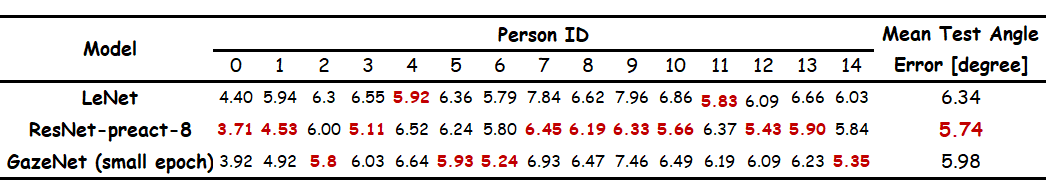
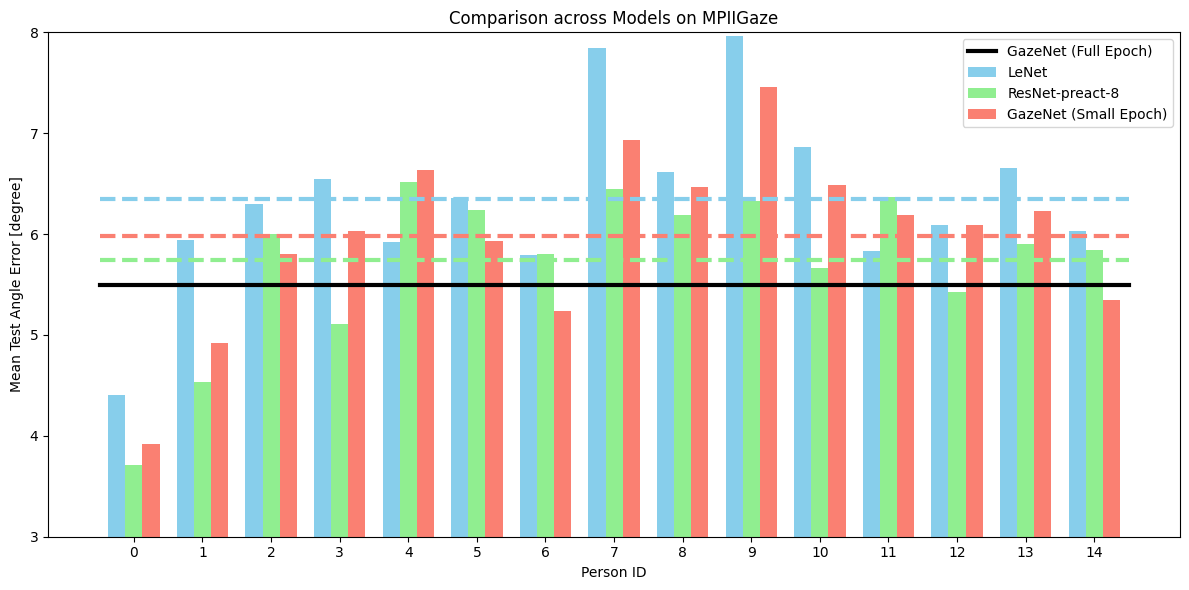
****

表1：在MPIIGaze上的实验结果（红色表示最优）

观察实验结果，三种模型在不同测试对象上的角度误差存在一定差异。其中，ResNet-preact-8 的平均测试角度误差最低（5.74°），表现出最优的整体性能，完全训练的GazeNet（full epoch）可以取得5.50°的平均误差，而GazeNet（small epoch）的平均误差为 5.98°，略高于 ResNet-preact-8，但仍优于 LeNet，这说明其在有限训练轮次下仍能取得较好的结果。相比之下，由于其浅层网络难以捕获复杂的眼动特征，LeNet 的平均误差最高（6.34°）。

****

图：实验结果柱状图(黑色来自[10])

在单个测试对象上，ResNet-preact-8 对大多数 Person ID 的误差较低，表现较为稳定（方差为0.56），而 LeNet （方差为0.68）和 GazeNet（small epoch）（方差为0.70） 在个别对象上的误差波动较大。排除训练不完全的影响，这一定程度表明深度模型在眼动跟踪任务中具有更高的鲁棒性。

**1.3 样例**

以下展示了resnet-preact在实际情况下的预测情况，样例1只有眼部移动，样例2增加了头部运动和眼镜。





附录

[1] Majaranta P, Bulling A. Eye tracking and eye-based human-computer interaction[M]//Advances in Physiological Computing. 2014: 39–65.

[2] Sugano Y, Zhang X, Bulling A. Aggregaze: Collective estimation of audience attention on public displays[C]//Proceedings of the ACM Symposium on User Interface Software and Technology. 2016: 821–831.

[3] Sattar H, Müller S, Fritz M, et al. Prediction of search targets from fixations in open-world settings[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 981–990.

[4] Funes Mora K A, Odobez J-M. Person independent 3D gaze estimation from remote RGB-D cameras[C]//Proceedings of the IEEE International Conference on Image Processing. 2013: 2787–2791.

[5] Schneider T, Schauerte B, Stiefelhagen R. Manifold alignment for person independent appearance-based gaze estimation[C]//Proceedings of the IEEE International Conference on Pattern Recognition. 2014: 1167–1172.

[6] Sugano Y, Matsushita Y, Sato Y. Learning-by-synthesis for appearance-based 3D gaze estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 1821–1828.

[7] Wood E, Baltrušaitis T, Morency L-P, et al. Learning an appearance-based gaze estimator from one million synthesised images[C]//Proceedings of the ACM Symposium on Eye Tracking Research and Applications. 2016: 131–138.

[8] Wood E, Bulling A. EyeTab: Model-based gaze estimation on unmodified tablet computers[C]//Proceedings of the ACM Symposium on Eye Tracking Research and Applications. 2014: 207–210.

[9] Zhang Y, Bulling A, Gellersen H. Sideways: A gaze interface for spontaneous interaction with situated displays[C]//Proceedings of the ACM CHI Conference on Human Factors in Computing Systems. 2013: 851–860.

[10] Zhang X, Sugano Y, Fritz M, Bulling A. MPIIGaze: Real-world dataset and deep appearance-based gaze estimation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 41(1): 162-175.

[11] Katrychuk D, Griffith HK, Komogortsev OV. Power-efficient and shift-robust eye-tracking sensor for portable VR headsets[C]//Proceedings of the 11th ACM symposium on eye tracking research and applications. New York: Association for Computing Machinery, 2019: 1-8. doi: 10.1145/3314111.3319821.

[12] de Lope J, Grana M. Deep transfer learning-based gaze tracking for behavioral activity recognition[J]. Neurocomputing, 2022, 500: 518-527. doi: 10.1016/j.neucom.2021.06.100.

[13] Bublea A, C˘aleanu CD. Deep learning based eye gaze tracking for automotive applications: An auto-keras approach[C]//Proceedings of the 2020 International Symposium on Electronics and Telecommunications (ISETC). Piscataway: IEEE, 2020: 1-4. doi: 10.1109/ISETC50328.2020.9301091.

[14] Wong ET, Yean S, Hu Q, et al. Gaze Estimation Using Residual Neural Network[C]//Proceedings of the 2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops). Piscataway: IEEE, 2019: 411-414. doi: 10.1109/PERCOMW.2019.8730846.

[15] Naqvi RA, Arsalan M, Batchuluun G, et al. Deep learning-based gaze detection system for automobile drivers using a NIR camera sensor[J]. Sensors, 2018, 18(2): 456. doi: 10.3390/s18020456.

[16] Illahi GK, Siekkinen M, Kamarainen T, et al. Real-time gaze prediction in virtual reality[C]//Proceedings of the 14th International Workshop on Immersive Mixed and Virtual Environment Systems. New York: Association for Computing Machinery, 2022: 12-18. doi: 10.1145/3534086.3534331.

[17] Ou WL, Kuo TL, Chang CC, et al. Deep-learning-based pupil center detection and tracking technology for visible-light wearable gaze tracking devices[J]. Applied Sciences, 2021, 11(2): 851. doi: 10.3390/app11020851.

[18] Chaudhary AK, Kothari R, Acharya M, et al. RITnet: Real-time semantic segmentation of the eye for gaze tracking[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Piscataway: IEEE, 2019: 3698-3702. doi: 10.1109/ICCVW.2019.00568.

[19] Kothari RS, Bailey RJ, Kanan C, et al. EllSeg-Gen, towards domain generalization for head-mounted eyetracking[J]. Proceedings of the ACM Human-Computer Interaction, 2022, 6: 139:1-139:17. doi: 10.1145/3530880.

[20] Chugh S. An eye tracking system for a virtual reality headset[D]. Ann Arbor: University of Toronto, 2020. ISBN: 9798698549024. [2024-10-10]. https://www.proquest.com/docview/2467610518/abstract/A2F4AD0C9231462CPQ/1.

[21] Lu C, Chakravarthula P, Liu K, et al. Neural 3D gaze: 3D pupil localization and gaze tracking based on anatomical eye model and neural refraction correction[C]//Proceedings of the 2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). Piscataway: IEEE, 2022: 375-383. doi: 10.1109/ISMAR558272022.00053.

[22] Gudi A, Li X, van Gemert J. Efficiency in real-time webcam gaze tracking[C]//Bartoli A, Fusiello A, eds. Computer Vision – ECCV 2020 Workshops. Cham: Springer International Publishing, 2020: 529-543. doi: 10.1007/978-3-030-66415-2.