

文章编号: 1002-1582(2026)04-0491-09

基于特征解耦的自监督视线估计方法

秦善彬, 陈靖*, 张湑琪, 符山沅

(北京理工大学 光电学院, 北京 100081)

摘 要: 视线估计作为眼动追踪的关键技术, 广泛应用于人机交互。现有方法一般依赖大量高精度注视标注面部图像, 然而标注这些数据需要大量人力与物力。工作提出了一种基于特征解耦的自监督视线估计方法, 根据视线估计预训练模型生成的伪标签构建注视方向相似图像对和身份与头姿相似图像对, 并计算其注视相似度, 采用注视解耦双编码器, 分别从相似图像对中提取与注视方向相关的特征, 以及头部姿态、身份信息或与注视无关的特征; 随后, 通过语义解耦判别模块预测不同类型特征间的语义相似性。同时使用交叉重构对该图像对的注视相似度迭代更新, 进而完成与注视方向相关与无关的特征的解耦。实验表明, 方法能有效将注视相关特征与无关特征解耦, 相较于现有方法取得了更优异的性能, 在MPIIFaceGaze数据集上的视线估计角度误差低至 4.06° 。

关 键 词: 视线估计; 自监督学习; 特征解耦; 人机交互; 表征学习

中图分类号: TP391.41

文献标识码: A

Self-supervised gaze estimation method based on feature disentanglement

QIN Shanbin, CHEN Jing*, ZHANG Tianqi, FU Shanfeng

(School of Optics and Photonics, Beijing Institute of Technology, Beijing 100081, China)

Abstract: Gaze estimation, as a key technology in eye movement tracking, is widely applied in human-computer interaction. Existing methods generally rely on a large number of facial images annotated with high-precision gaze directions, yet annotating such data incurs substantial human and material costs. This work proposes a self-supervised gaze estimation method based on feature disentanglement, which constructs gaze direction similar image pairs and identity & head pose similar image pairs using pseudo-labels generated by a pre-trained gaze estimation model, and computes their gaze similarity. A gaze disentanglement dual encoder is employed to extract gaze direction-related features and gaze-irrelevant features (such as head pose and identity information) from similar image pairs, respectively. A semantic disentanglement discriminative module then predicts the semantic similarity between different types of features. Meanwhile, cross-entropy reconstruction is used to iteratively update the gaze similarity of these image pairs, thereby achieving the disentanglement of gaze direction-related and gaze-irrelevant features. Experiments demonstrate that the proposed method can effectively disentangle gaze-related features from irrelevant features and achieves superior performance compared to state-of-the-art methods, with a gaze estimation angular error as low as 4.06° on the MPIIFaceGaze dataset.

Keywords: gaze estimation; self-supervised learning; feature disentanglement; human-computer interaction; representation learning

0 引 言

视觉作为人类获取外部信息的核心感官通道,

其信息筛选过程与注视行为高度关联。注视方向不仅直观反映了人类对周围环境的感知优先级与注意力分配模式, 更成为解析人机交互意图、环境认知

收稿日期: 2026-02-11; 收到修改稿日期: 2026-03-01

基金项目: 航空科学基金 (2022Z005072001)

作者简介: 秦善彬 (2001—), 男, 硕士研究生, 从事视线估计、深度学习研究。3120230526@bit.edu.cn

陈靖 (1974—), 女, 教授, 博士, 从事虚拟现实与增强现实, 视觉SLAM, 深度学习研究。

* 通讯作者: chen74jing29@bit.edu.cn

机制的关键切入点。因此,视线估计技术已在虚拟现实^[1-2]、智能驾驶辅助系统^[3-4]、认知心理学实验研究^[5]等多个前沿领域广泛应用,成为当前计算机视觉与人机工程交叉领域的研究热点之一。

近年来,基于外观的注视估计凭借非侵入式、适配复杂场景的优势成为主流方向。此类方法多采用有监督框架,通过卷积神经网络提取面部或眼部图像中与注视相关的关键视觉特征,经多层感知器端到端回归,输出俯仰角与偏航角表征的二维注视向量。但有监督方法的精度高度依赖高质量标注数据,其真值获取依托高精度眼动追踪设备,导致大规模数据集构建成本高。为突破这一依赖,现有方法通常利用面部图像变换中的注视方向一致性进行自监督约束,进而完成视线估计。Yu^[6]等通过建模输入眼图对间的视线变换关系来学习鲁棒的注视表征,设计了注视表征网络与重定向网络的联合训练机制。该方法利用注视表征网络分别从输入眼图与目标眼图中提取注视特征,将二者的差异输入至重定向网络,以生成与目标眼图注视方向一致的输出眼图,同时通过图像域的损失函数约束其生成的准确性。Gu^[7]等对输入面部图像进行数据增强以构建增强图像对,并将其送入编码器,通过不变性、相似性、方差和协方差项损失函数,约束编码器两分支输出隐向量的一致性,从而实现注视特征的有效提取。SwAT^[8]提出一种基于数据增强交换的自监督注视估计方法,对面部图像应用包含外观与几何变换并生成两个增强视图,利用特征变换层交换两视图的几何变换信息并映射至特征空间,再通过在线聚类损失约束变换后的特征一致性,从而学习到与几何变换等变的注视表征。然而,这类方法仅考虑图像变换中注视方向的一致性,未排除光照变化、头姿偏移、面部遮挡及外观信息等干扰因素,此类注视无关特征与注视特征耦合复杂、难以人工分离,进而影响注视回归精度。

为了解决上述问题,研究者们尝试将注视干扰因素分离。为了将注视特征与眼部外观特征分离,Sun^[9]等提出了一种基于交叉重构的自监督方法,以眼部外观相似眼图对与视线相似眼图对为输入,将眼图对分别编码为注视特征与眼部外观特征,通过解码器将交换后的特征融合来重构眼图,使用 $L1$ 损失函数约束重构眼图与输入眼图的一致性,实现注视特征和眼部外观特征解耦,提取更精确的注视表征。Ni^[10]等设计了特征投影模块,将面部图像特征分解为注视特征与头部姿态特征。然后将该模块提取的面部注视特征与眼部图像提取的注视特征融合以捕捉眼部外观的细微变化,增强注视特征表

征,同时将头姿特征用于注视估计的辅助校正。尽管上述方法实现了注视特征与注视无关特征的解耦,但仅考虑了头部姿态、外观信息等单一注视干扰因素,解耦效果仍有不足。为消除更多类注视干扰,本文针对面部图像中注视相关特征与身份、头姿等干扰特征高度耦合的问题,以伪标签引导的相似图像对构建为基础,设计注视解耦双编码器、语义解耦判别模块、交叉与重构模块的协同训练,同时通过迭代优化策略自适应伪标签噪声,提升注视特征提取的鲁棒性与准确性。

本方法的主要贡献如下:

(1) 提出伪标签引导的双分支特征解耦自监督学习方法,通过构建注视相似图像对和身份头姿相似图像对为判别器提供监督信号,设计双编码器与判别器协同架构,实现注视特征与身份、头姿等干扰特征的显式分离,有效提升了注视特征的纯度。

(2) 提出基于交叉重构的相似度迭代优化方法,通过计算交叉重构的结果与输入面部图像的结构相似度,自适应修正输入图像对的注视相似度,有效缓解了伪标签噪声带来的监督偏差,进一步提升了特征解耦的稳定性。

(3) 大量实验证明了所提方法的有效性和可行性,在MPIIFaceGaze数据集上的精度误差可达 4.06° ,相较于先进视线估计方法CA-Net精度提升 4.9% 。

1 相关工作

1.1 有监督视线估计

近年来,基于深度学习的方法已成为注视估计领域的主流趋势。基于外观的注视估计方法以眼部图像或单张面部图像作为网络输入,直接输出注视方向。早期的监督方法^[14]将提取的特征与头部姿态真值进行拼接,但未能充分利用全脸信息,造成信息浪费。为了充分利用面部信息,Zhang^[11]等提出了一种以全脸图像为输入的视线估计方法,并通过空间权重增强面部特定区域特征,进而辅助视线估计任务。此后,众多研究开始采用面部图像作为网络输入。为了无约束环境下的视线估计泛化,Abdelrahman^[12]等分别对注视方向的俯仰角和偏航角设计损失函数并预测。为了捕捉由注视方向变化引起的面部和眼部外观细微变化,Chen^[13]等将膨胀卷积引入了视线估计领域,通过扩大卷积感受野的同时保持特征图分辨率不变,整合眼部细节与面部全局上下文信息,提升了模型鲁棒性。为捕捉眼部区域精细的特征,Abdelrahman^[14]等通过引入渐进式的注意融合机制,通过使用旋转矩阵形式的注视标签,解决了球面角表示

不连续与模糊的问题,进而提升视线回归精度。为了解决面部图像低分辨率、黑暗与遮挡的问题,Huang^[15]等通过从变分隐S空间采样生成多个眼球掩模并进行注视回归,以基于真值注视构建的眼球掩模为监督信号,引导网络聚焦眼部关键注视信息。然而,有监督视线估计需要大量的带有真值标签的人脸图像,不仅耗时,而且需要较为复杂的设备支持。因此,视线估计逐渐转向了自监督训练方式。

1.2 自监督视线估计

自监督表征学习的核心目标是学习高质量的特征,为下游任务提供有效支撑。在缺乏注视标注的场景下,学习纯净的注视特征至关重要^[16-17]。早期研究主要以眼部图像作为网络输入。Mahmud^[18]等对眼部区域分割网络 and 对比学习网络进行预训练,其中前者生成眼部角膜与瞳孔的掩码,后者学习面部裁剪所得成对眼部图像的注视特征一致性,最后将眼部图像和生成的掩码输入到对比网络提取注视特征,并将其用于注视方向回归。为了削弱身份信息对视线估计的干扰,Du^[19]等通过

保留眼部区域的裁剪与色彩畸变,构建注视一致图像对与注视对比图像对并送入编码器,同时引入受试者条件投影模块,将编码器的输出映射至多个受试者专属嵌入空间,再通过对比损失函数对编码器输出进行约束,最终使编码器学习到兼具注视感知性与通用性的表征。Sun^[9]等以同一眼部的外观相似图像对和同一帧双眼的视线相似图像对为训练输入,将每张图像编码为相互独立的注视特征与眼部身份特征;通过特征交换与交叉重构机制,约束注视特征仅编码视线方向信息、眼部身份特征仅捕捉眼部外观与身份属性,最终学习到兼具互斥性与表征完整性的解耦特征。然而,这些方法未能考虑或只考虑了单一注视干扰因素对视线估计的影响。为此,本文提出了一种能同时分离出身份信息与头部姿态的注视特征解耦方法。

2 基于特征解耦的注视估计方法

本节提出的基于特征解耦的自监督视线估计方法的架构图如图1所示。该方法首先利用预训练

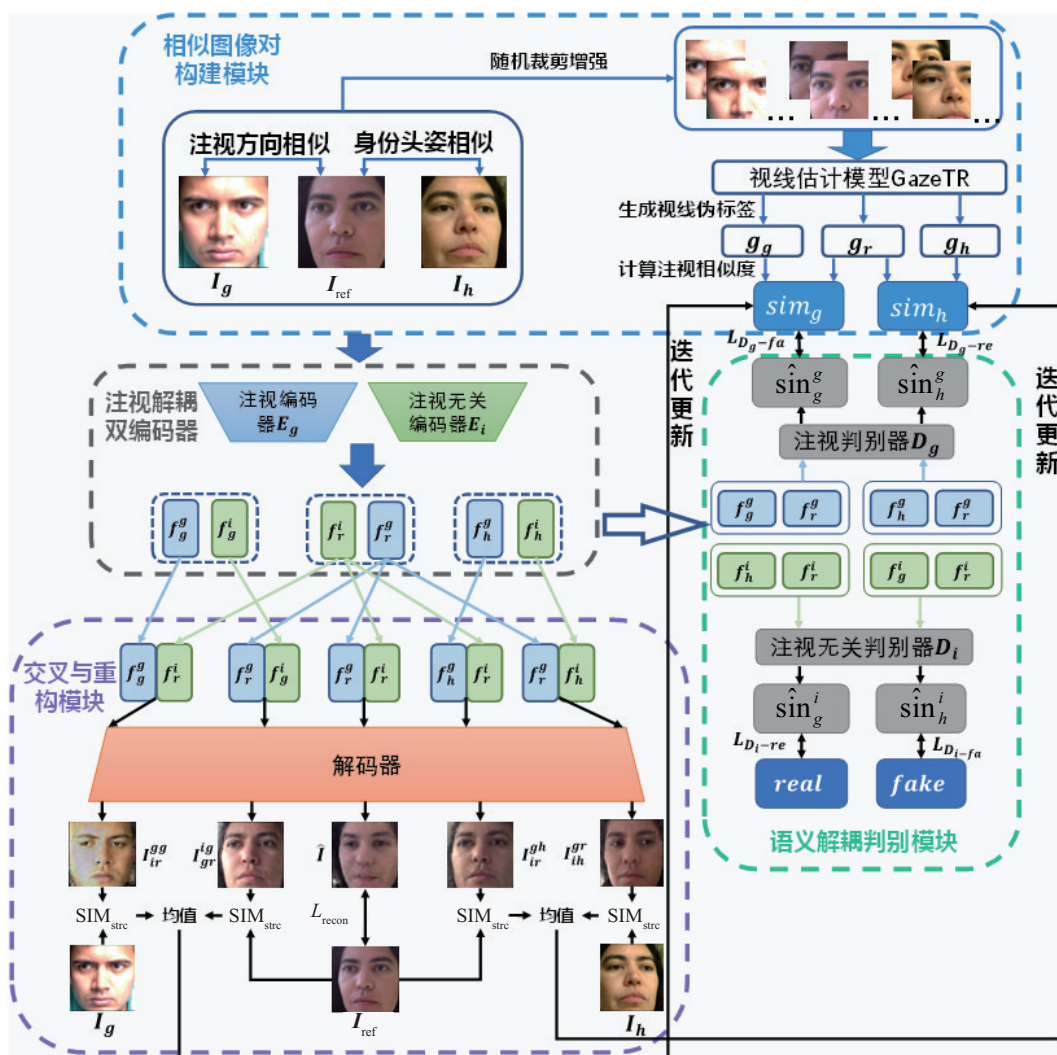


图1 基于特征解耦的自监督注视估计方法框架图

模型（例如 GazeTR^[20]）生成的伪标签，构建注视方向相似和身份与头姿相似两类面部图像对，并计算初始注视相似度；通过注视解耦双编码器分别对图像对进行注视相关与无关特征的编码，利用语义解耦判别模块完成同类型特征的语义相似性判别。在此过程中，通过计算交叉重构结果与输入面部图像的结构相似性，迭代更新面部图像对的注视相似度，实现对伪标签噪声的自适应调整，进而完成注视特征的提取与分离。

2.1 相似图像对构建模块

为了向语义解耦判别模块提供监督信号，本文设计了注视相似面部图像对和身份与头姿相似面部图像对作为网络的输入。本文使用预训练视线估计模型为训练集数据生成伪标签并计算面部图像对间的注视相似度，进而完成相似图像对的构建。

设训练集 R 包含 n 个独立个体样本，记为 $R_1, R_2 \dots R_n$ 。每个个体样本 R_i 包含若干张面部图像，对其中任一面部图像 I_{ref} ，本文利用在一个公开的视线估计数据集 S 上训练的视线估计模型来生成伪标签 g_{ref} ，同时对 I_{ref} 进行 5 次随机裁剪与增强后再进行视线估计。为衡量预训练模型对训练集的预测稳定度，通过这 5 次估计结果的均值与方差计算伪标签置信度 conf_r

$$v_{\text{total}} = w_p v_p + w_y v_y \quad (1)$$

$$\text{conf}_r = \exp\left(-\frac{v_{\text{total}}}{v_{\text{max}}}\right) \quad (2)$$

式中， v_p, v_y 分别为俯仰、偏航角的方差； w_p, w_y 分别为俯仰、偏航角的方差权重； v_{max} 为训练数据的最大 v_{total} 。同时使用在一个公开的头部姿态数据集 T 上训练的头部姿态估计模型生成伪标签 h_{ref} 。随后，对每个个体样本 R_i 的每张面部图像 I_{ref} ，从 R 中排除 R_i 后的剩余样本中，筛选出注视伪标签与 I_{ref} 最接近且 $\text{conf}_r \geq 0.8$ 的图像 I_g ，并将二者构建为注视相似图像对，以保证图像对的样本对象身份信息的差异，同时保证二者注视方向的相似。本文定义注视相似度以表征一对面部图像的注视方向相似性。对于俯仰角和偏航角两个维度，注视相似度的计算方式如下

$$\text{sim}_{\text{mean}} = \exp\left(-\frac{(m_1 - m_2)^2}{2\sigma^2}\right) \quad (3)$$

$$\text{sim}_{\text{var}} = \frac{2\min(v_1, v_2)}{\max(v_1, v_2) + \min(v_1, v_2)} \quad (4)$$

$$\text{sim} = w_{\text{mean}} \text{sim}_{\text{mean}} + w_{\text{var}} \text{sim}_{\text{var}} \quad (5)$$

式中， m_1, m_2 为该图像对预测注视方向的均值； v_1, v_2 为方差； $w_{\text{mean}}, w_{\text{var}}$ 分别为均值相似度 sim_{mean} 与方差相似度 sim_{var} 的权重，那么总注视相似度可表示为

$$\text{sim}_g = \text{sim}_{\text{pitch}} + \text{sim}_{\text{yaw}} \quad (6)$$

式中 $\text{sim}_{\text{pitch}}, \text{sim}_{\text{yaw}}$ 分别为俯仰、偏航方向的相似度。

身份与头姿相似图像对的构建遵循类似逻辑。对于训练集 R 中每个独立样本 R_i ，以其面部图像 I_{ref} 为基准，在 R 内筛选头部姿态标签最为接近的另一图像 I_h ，以此构成身份与头姿相似图像对，确保该图像对的身份和头部姿态信息高度一致。随后，可依据式 (3) ~ (6) 计算该图像对的注视方向相似度 sim_h 。

2.2 注视解耦双编码器

在构建了相似图像对的基础上，为分离注视特征与身份、头姿特征，本文设计了注视解耦双编码器，分别对应注视编码器 E_g 与身份及头姿编码器 E_i 。视线估计任务不仅需要对眼部区域的瞳孔位置建模，还要建模整个面部图与眼部的复杂关系，这就要求编码器应当具备全局与局部的感知能力。因此，本文使用 ResNet 并在每个 Block 中添加了 1×1 卷积、 1×1 卷积 + 3×3 卷积、 3×3 卷积 + 3×3 池化等多个分支以捕捉全局信息与眼部区域信息。对于输入面部图 I_{ref} ，注视相关特征 f_g^r 和无关因素的特征 f_i^r 可以由下式得到

$$f_g^g = E_g(I_{\text{ref}}) \quad (7)$$

$$f_i^i = E_i(I_{\text{ref}}) \quad (8)$$

添加多分支导致了模型参数增加，显存占用提升。为了应对显存增加的问题，本文采用一种受 OREPA^[21] 启发的在线卷积重参数化策略，该方法将多分支模块线性化，以逐通道线性缩放层替代非线性 BN 层，保持分支优化多样性且支持在线等价融合；在训练阶段将多分支、多层卷积压缩为单一卷积，把特征图上的高开销运算转为卷积核运算，大幅降低显存与计算消耗。推理时等价于普通卷积，不引入额外参数数量与计算量，在保证模型性能与部署效率的同时，显著提升训练效率。

2.3 语义解耦判别模块

为使双编码器输出的特征互斥，本文引入语义解耦判别模块，通过判别器对同类型特征相似性的学习，引导编码器将注视特征和身份与头姿特征分离，进而实现注视特征的纯化。以注视相似图像对为例，判别器需准确识别该图像对的注视特征的语义相似性，并区分注视无关特征的语义差异性。

为了不同类型特征的判别，本文构建了两个判别器，注视判别器 D_g 和注视无关判别器 D_i ，用于判断面部图像经编码后特征之间的相似性。将 E_g 对 I_{ref} 的编码结果记为 f_g^g ， E_i 对 I_{ref} 的编码结果记为 f_i^i 。类似地，本文统一符号规范：下标表示输入图像的来源，上标表示编码器的类型。其中，下

标 g 、 h 、 r 分别对应输入图像 I_g 、 I_h 与 I_{ref} ; 上标 g 代表注视编码器 E_g 输出的注视相关特征, i 代表 E_i 输出的特征。

对于注视相似图像对, 将 f_r^g 和 f_g^g 送入 D_g 与 D_i , 输出注视相似性分数

$$\hat{\text{sim}}_g^g = D_g(f_r^g, f_g^g) \quad (9)$$

$$\hat{\text{sim}}_g^i = D_i(f_h^i, f_r^i) \quad (10)$$

同理, 对于头部姿态相似图像对有

$$\hat{\text{sim}}_h^g = D_g(f_r^g, f_h^g) \quad (11)$$

$$\hat{\text{sim}}_h^i = D_i(f_r^i, f_g^i) \quad (12)$$

2.4 交叉与重构模块

为约束注视解耦双编码器输出的特征, 本文通过将 I_{ref} 编码后的的注视特征与注视无关特征一同送入解码器执行图像重构任务

$$\hat{I} = D(f_r^g, f_r^i) \quad (13)$$

式中, $\hat{I}$ 为重构的面部图像; D 为一个解码器, 负责将输入的注视特征与注视无关特征拼接, 并重构为与输入尺寸一致的面部图像。

然而, 2.1 节中面部图像的注视相似度计算仍存在可靠性不足的问题。为使模型在含噪伪标签中获取有效监督、提升注视相似度准确性并增强特征解耦效果, 本文提出基于交叉重构的迭代优化策略, 其结果可表示为

$$I_{ir}^{gg} = D(f_r^i, f_g^g) \quad (14)$$

$$I_{gr}^{ig} = D(f_r^g, f_g^i) \quad (15)$$

$$I_{ih}^{gr} = D(f_r^g, f_h^i) \quad (16)$$

$$I_{ir}^{gh} = D(f_r^i, f_h^g) \quad (17)$$

如果该相似图像对注视或身份与头姿相似, 那么其特征的交叉重构将会保留彼此特征的相似部分, 重构结果将与输入图像对相似。为量化面部图像对的交叉重构效果, 本文定义注视相似对重构相似度 α_g 和身份头姿相似对重构相似度 α_h , 二者基于结构相似性指数计算

$$\text{SIM}_{\text{strc}}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (18)$$

式中, μ_x, μ_y 为图像均值; σ_x, σ_y 为图像方差, σ_{xy} 为协方差; C_1, C_2 为常数。为了提升注视相似度的准确性, 本文使用交叉重构的结果对输入面部图像对的注视相似度进行迭代更新

$$\text{sim}_t = \alpha_g \text{sim}_{t-1} + (1 - \alpha_g) \text{sim}_{t-2} \quad (19)$$

式中, $\alpha_g = \frac{\text{SIM}_{\text{strc}}(I_{ir}^{gg}, I_{\text{ref}}) + \text{SIM}_{\text{strc}}(I_{gr}^{ig}, I_g)}{2}$, $\alpha_h = \frac{\text{SIM}_{\text{strc}}(I_{ih}^{gr}, I_{\text{ref}}) + \text{SIM}_{\text{strc}}(I_{ir}^{gh}, I_g)}{2}$ 分别表示第 t 轮迭代时, 相似图像对经交叉重构后, 重构结果与输入

图像对的结构相似性均值, 它衡量了两幅图像在亮度、对比度及结构一致性方面相似程度, 其取值范围为 $[0, 1]$, 数值越接近 1 表明图像相似性越高。

2.5 损失函数

考虑到既要提取完整的面部特征, 又要针对性地对注视特征与注视无关特征解耦, 本文设计了两种损失函数, 模型训练使用二者的线性组合

$$L = L_{\text{recon}} + \lambda L_D \quad (20)$$

式中 λ 表示用于平衡权重的系数; L_{recon} 为面部图像重构损失; L_D 为判别器损失。下面对各损失函数说明。

2.5.1 重构损失

为了提取到有效的面部图像特征, 本文首先设计了重构任务, 采用 $L1$ 损失作为图像重构约束

$$L_{\text{recon}} = L_1(I_{\text{ref}}, \hat{I}) \quad (21)$$

2.5.2 判别损失

对于判别器, 损失函数采用稳定的二元交叉熵损失

$$L_{\text{sBCE}}(p, l) = -\frac{1}{N} \sum_{i=1}^N [\alpha l_i \cdot \log(\max(p_i, \epsilon)) + (1 - \alpha)(1 - l_i) \cdot \log(\max(1 - p_i, \epsilon))] \quad (22)$$

式中 α 表示标签平滑系数; ϵ 表示数值稳定性参数; N 表示批大小。由此, 得到判别器 D_g 的损失函数

$$L_{D_g-re} = L_{\text{sBCE}}(\hat{\text{sim}}_g^g, \text{sim}_g) \quad (23)$$

$$L_{D_g-fa} = L_{\text{sBCE}}(\hat{\text{sim}}_h^g, \text{sim}_h) \quad (24)$$

式中, sim_g 表示注视相似图像对的注视相似度, sim_h 表示身份与头姿相似图像对的注视相似度。类似地, 注视无关判别器 D_i 的损失函数可表示为

$$L_{D_i-re} = L_{\text{sBCE}}(\hat{\text{sim}}_g^i, l_r) \quad (25)$$

$$L_{D_i-fa} = L_{\text{sBCE}}(\hat{\text{sim}}_h^i, l_f) \quad (26)$$

式中, l_r 表示真实标签, l_f 表示虚假标签, 分别表示身份与头姿相似图像对的注视无关特征相似与否。最终, 判别器的训练目标为最小化如下损失

$$L_D = \frac{1}{4} (L_{D_g-re} + L_{D_g-fa} + L_{D_i-re} + L_{D_i-fa}) \quad (27)$$

3 实验

3.1 实验设置

3.1.1 数据集

为验证本文所提方法的有效性, 本文在三个主流注视估计数据集 MPIIFaceGaze^[22]、EyeDiap^[23] 与 ETH-XGaze^[24] 上开展实验。

MPIIFaceGaze 数据集包含 15 名受试者的 212659 张面部图像。为保证数据一致性, 本文遵循文献^[25]的预处理流程, 依次完成面部关键点对齐与眼部区域裁剪。在实验阶段, 使用留一法交叉

验证进行评估。

EyeDiap数据集包含原始数据为16名受试者的94个视频片段,本文参照文献[25]的筛选标准,从14名受试者的视频中提取有效帧,从每名对象中选取1100张标准化面部图像,最终构建包含1.54万张图像的实验数据集。在实验阶段,采用与文献^[26]一致的5折交叉验证方案进行评估。

ETH-XGaze数据集包含80名受试者的75.6万

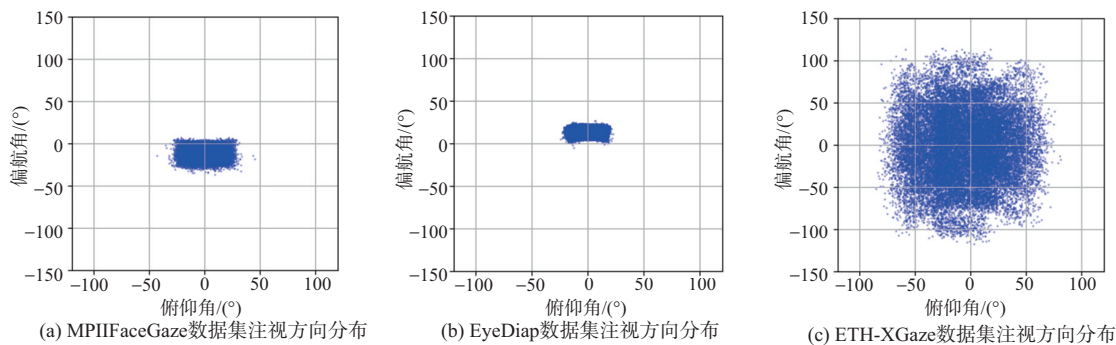


图2 MPII数据集、EyeDiap数据集与ETH-XGaze数据集的注视方向分布函数

3.1.2 实现细节

本文方法基于PyTorch 2.4.1实现,在NVIDIA GeForce RTX 3090进行了实验,并基于CUDA 12.2平台加速计算。使用的操作系统为Ubuntu 20.04,开发语言为Python 3.8.20。

本文方法中的解码器采用4个反卷积块^[26]构成,判别器由4个交替的卷积-ReLU层构成。本文使用在ETH-XGaze数据集上训练的GazeTR模型^[20](训练50个epoch,批次大小为512)为MPIIFaceGaze数据集生成伪标签;使用在MPIIFaceGaze数据集上训练的Full-Face模型^[26](训练40个epoch,批量大小为32)为EyeDiap数据集生成伪标签。两个数据集的头部姿态伪标签均使用TokenHPE模型^[27]生成,该模型是一种关联面部部件的Transformer架构模型,属于当前头部姿态估计领域的先进方法之一。输入图像通过线性缩放调整为 224×224 大小。训练过程采用Adam^[28]优化器,在MPIIFaceGaze、EyeDiap和ETH-XGaze数据集上的训练epoch数分别为70、110和50,解码器,判别器和编码器的学习率分别为0.0005、0.0005和0.0002。

3.2 性能评估

3.2.1 数据集内评估

本文将所提方法与近年来与自监督视线估计领域的代表性方法进行比较,包括对比学习方法SimCLR^[16]、BYOL^[17],贝叶斯对抗方法^[29]以及Cross-Encoder方法^[9]。为遵循自监督实验设置,所有方法均首先在无标注数据上训练双编码器,然后利用随机选取的200个标注样本,通过一个轻量

张图像。本文选取其中15名受试者的数据作为测试集,剩余65名受试者的数据作为训练集。预处理仅进行基础的面部对齐与图像尺寸归一化。

为直观展示各数据集的注视分布特征,本文对上述三个数据集的注视方向(以俯仰角和偏航角表征)进行了统计分析,其可视化结果如图2所示。可以看出,不同数据集在注视角度覆盖范围与分布密度上存在明显差异。

级MLP注视回归器对编码器进行微调。表1展示了各方法在MPIIFaceGaze、EyeDiap数据集上的平均注视角度误差对比结果。可以看到,本文方法在2个数据集上均取得了最低的注视误差,表明基于特征解耦的设计可有效提取与注视相关的判别性特征。相比之下,SimCLR^[16]和BYOL^[17]等通用自监督框架因难以捕捉眼部区域的细微变化,性能表现较弱。虽然本文的数据配对方法与Cross-Encoder^[9]有相似之处,但是Cross-Encoder只解耦出了眼部的外观特征,未能分离头部姿态与身份信息,因此注视精度较低,而本文方法的注视误差较之降低了29.4%。与引入贝叶斯对抗的方法^[29]相比,本文方法亦将注视误差进一步降低12.7%。

为评估所提方法在标注数据充足场景下的性能上限,本文进一步将其与多种有监督方法进行了对比,包括GazeNet^[30]、FullFace^[15]、Deep Pictorial^[31]、AR-Net^[32]、Dilated-Net^[13]、CA-Net^[33]、GazeTR-Pure^[20]、FARE-Net^[34]、ETH-XGaze^[24]和LRSNet^[35]。评估过程中,在EyeDiap数据集上进行测试时,首先在MPIIFaceGaze数据集上使用本文方法进行预训练,再使用EyeDiap的全部训练数据对编码器与注视回归器进行微调;在MPIIFaceGaze数据集上测试时同理。表1展示了有监督对比实验结果(部分角度误差数据来自文献[25]和相关原始论文)。结果表明,在数据量充足时,本文方法仍优于主流的有监督视线估计方法,进一步验证了特征解耦策略的有效性。

表 1 数据集内注视角度误差实验对比结果(单位:度)

监督类型	方法	MPII	EyeDiap
自监督	SimCLR ^[16]	9.64	10.23
	BYOL ^[17]	9.41	10.58
	Cross-Encoder ^[9]	7.3	N/A
	Bayesian ^[29]	5.9	9.9
	Ours	5.15	8.29
有监督	GazeNet ^[30]	5.76	6.79
	FullFace ^[15]	4.93	6.53
	Pictorial ^[31]	4.5	10.3
	AR-Net ^[32]	5.65	6.35
	Dilated-Net ^[13]	4.42	6.19
	CA-Net ^[33]	4.27	5.27
	GazeTR-Pure ^[20]	4.74	5.72
	FARE-Net ^[34]	4.4	5.9
	ETH-XGaze ^[24]	4.8	6.5
	LRSNet ^[35]	4.5	6.6
	Ours	4.06	5.80

3.2.2 跨数据集评估

为测试本文模型的泛化能力,本实验采用跨数据集设定进行性能评估。具体而言,仅使用目标数据集的 100 个样本对预训练模型进行微调,以测试其在数据分布变化下的适应能力。表 2 展示了跨数据集评估结果。在本实验中,ETH-XGaze (X)、MPIIFaceGaze (M) 和 EyeDiap (E) 被用于三种交叉验证设定。在跨数据集评估中,本文方法在 E-M 与 X-M 设置上均取得最优结果,仅在 M-E 设置下误差略高于 Cross-Encoder,这一现象源于两种方法的解耦对象与输入空间存在本质差异:Cross-Encoder 仅以眼部图像为输入、仅解耦注视与眼部外观两类特征,输入区域小、干扰因素少,且在 EyeDiap 这类眼部分辨率较低的数据集上,局部眼部输入更易保持特征一致性,跨域适配速度更快;而本文方法以全脸图像为输入,同时解耦注视、身份、头部姿态三类强耦合特征,解耦难度更高、约束更强,在 M-E 这种域偏移剧烈、标注样本极少的跨域场景下,全脸分布差异会带来额外适配负担,但在 E-M 等域偏移更温和、数据更规范的场景中,强解耦能充分滤除干扰,性能优势显著。

表 2 跨数据集注视角度误差实验对比结果(单位:度)

方法	M-E	E-M	X-M
SimCLR ^[16]	10.76	7.67	11.07
BYOL ^[17]	10.24	7.26	11.34
Cross-Encoder ^[9]	8.32	6.42	9.04
Ours	8.46	6.34	8.07

其中,M、E、X 分别代表 MPII、EyeDiap、ETH-XGaze 数据集。M-E 表示在 MPII 进行预训练,在 EyeDiap 数据集进行微调。

3.3 消融实验

3.3.1 注视相似度迭代策略消融

本文设计基于交叉重构的自适应标签迭代更新策略,以抑制标签噪声、优化判别器监督信号。为验证该机制有效性,消融实验将其替换为仅基于原始伪标签构建相似图像对的训练方式,不计算和更新标签置信度与注视相似度,结果如表 3 所示。移除该策略后注视估计精度下降,证明该机制可有效抑制伪标签噪声,为判别器提供更可靠的监督信息。

表 3 注视相似度迭代更新策略消融实验的
注视角度误差对比(单位:度)

迭代更新策略	MPII	EyeDiap
×	5.30	8.54
✓	5.15	8.29

3.3.2 解耦消融

本文使用注视相似图像对和身份与头姿相似图像对用于训练判别器以实现注视相关特征与无关特征的解耦。为了验证这两类图像对输入的有效性,本节设计消融实验,依次仅保留其中一种相似图像对作为输入,并移除另一图像对对应的训练模块与损失函数。微调过程遵循自监督设定,使用 200 个标注样本,并在数据集内进行评估。实验设置中,仅使用注视相似对输入的配置记为 GS,仅使用头部姿态相似对输入的配置记为 PS。消融实验结果如表 4 所示。

表 4 输入图像对消融实验的注视角度误差对比(单位:度)

GS 对	HS 对	MPII	EyeDiap
×	✓	9.78	12.01
✓	×	7.68	10.52
✓	✓	5.15	8.29

可以看出,移除注视相似对后,模型性能显著下降,这主要是由于训练过程中缺乏对注视相关特征的显式关注与学习,导致编码器与判别器仅聚焦于头部姿态与身份信息。而去除了头部姿态相似对的模型性能也有所下降,因为仅依靠注视特征编码的学习难以将无关特征解耦出来。

3.3.3 损失函数权重系数消融

如式(20)所示,损失函数由重构损失 L_{recon} 与判别损失 L_D 两部分构成,并由权重系数 λ 调节。为了验证语义解耦判别模块的有效性,本节将 λ 分别设为 0、0.1、1、10 开展实验,结果如表 5 所示。可以看出, λ 取 0,即完全消除判别模块后,视线估计精度急剧下降,这主要是因为模型只学习了脸部特征,而未能学到注视特征。当 λ 取值由 1 逐

渐变增大时，判别模块逐渐占据主导，重构模块边缘化，对面部特征的处理能力被削弱，重构效果变差，进而影响对伪标签的更新过程，因此模型的估计精度下降。

表 5 损失函数权重系数 λ 消融
实验的注视角度误差对比(单位:度)

λ	MPII	EyeDiap
0	12.36	14.51
0.1	5.15	8.29
1	7.23	9.68
10	7.92	9.87

3.4 可视化分析实验

3.4.1 t-SNE可视化

为验证特征解耦的有效性，本文采用t-SNE^[36]对编码器输出的注视相关特征与注视无关特征进行可视化分析。由于两类特征的编码维度不一致，本文首先通过主成分分析将两者统一降至10维。可视化过程使用了MPIIFaceGaze和EyeDiap数据集的所有数据。图3(a)，(b)展示未经训练的编码器的输出特征分布，图3(c)，(d)展示训练后编码器的输出特征分布。对比可见，训练前的特征分布较为散乱，而经本文方法训练后，注视相关特征(蓝色)呈现出明显的聚类现象，且与注视无关特征(红色)之间界限清晰，这说明了本文所提方法能够有效实现注视相关特征与无关特征的解耦。

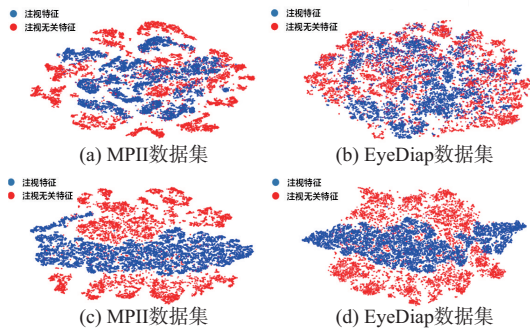


图 3 t-SNE 特征解耦可视化

3.4.2 重建可视化

为验证注视相关特征与注视无关特征的有效解耦，本文基于编码-解码器结构进行面部图像重建实验。具体而言，将提取到的两类特征分别置零并输入解码器进行重建，结果如图4所示。结果表明，将注视无关特征置零会破坏面部外观，但能保留注视方向；将注视相关特征置零则会恢复面部外观，但注视方向变得随机。这一结果证实了能够实现两类特征的解耦，且注视无关特征主要负责编码对象的面部外观信息。

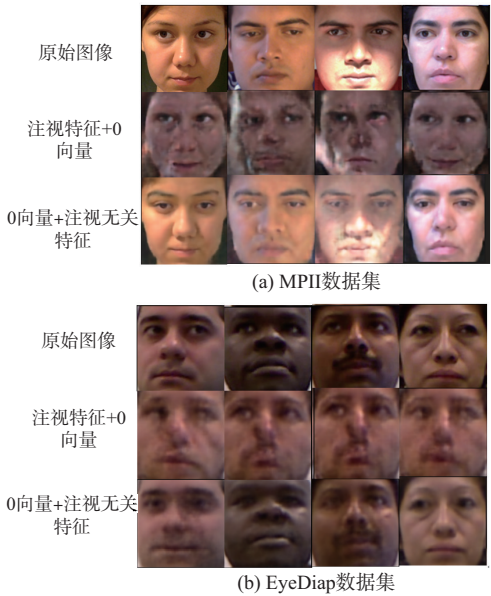


图 4 人脸重建效果

4 结 论

本文提出面向视线估计的自监督特征解耦方法，通过注视解耦双编码器、语义解耦判别模块与交叉重构模块的协同训练，解决了视线估计中标注数据不足的核心问题。该方法以预训练模型生成的伪标签为基础，构建两类相似图像对并迭代优化注视相似度，经双编码器编码、判别器语义判别与解码器重构约束，实现注视特征与身份、头姿等干扰特征的有效解耦，学习到互斥且完整的特征表征。实验表明，该方法在少样本和跨数据集评估中均优于现有方法，可视化与重建实验验证了其稳定的特征解耦效果。该方法性能依赖预训练模型的域泛化能力，伪标签质量直接影响估计精度，针对特定场景选择或域适应预训练模型，是进一步提升性能的关键。

参考文献:

[1] Clay V, König P, Koenig S. Eye tracking in virtual reality[J]. Journal of Eye Movement Research, 2019, 12(1): 3.

[2] Kouliris G A, Akşit K, Stengel M, et al. Near-eye display and tracking technologies for virtual and augmented reality[C]//Computer Graphics Forum. Vienna, Austria: Wiley, 2019: 493—519.

[3] Wang K, Gao Y, Song W, et al. Gaze pattern genius: gaze-driven vr interaction using unsupervised domain adaption[C]//2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW). Wellington, New Zealand:IEEE, 2024: 937—938.K.

[4] Swan G, Goldstein R B, Savage S W, et al. Automatic processing of gaze movements to quantify gaze scanning behaviors in a driving simulator[J]. Behavior Research Methods, 2021, 53(2): 487—506.

[5] Chuang M H, Chen C L, Ma J P. Eye tracking analysis of readers' psychological interaction with marketing copy referenc-

- ing life values[C]//International Conference on Human-Computer Interaction. Los Angeles, CA, USA: Springer International Publishing, 2015: 87—92.
- [6] Yu Y, Odobez J M. Unsupervised representation learning for gaze estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, Washington, USA: IEEE, 2022: 7314—7324.
 - [7] Gu D, Lv M, Liu J. Vicsgaze: a gaze estimation method using self-supervised contrastive learning[J]. *Multimedia Systems*, 2024, 30(6): 330.
 - [8] Farkhondeh A, Palmero C, Scardapane S, et al. Towards self-supervised gaze estimation[C]//33rd British Machine Vision Conference. London, UK: British Machine Vision Association, 2022: 1—13.
 - [9] Sun Y, Zeng J, Shan S, et al. Cross-encoder for unsupervised gaze representation learning[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, QC, Canada: IEEE, 2021: 3702—3711.
 - [10] Ni K, Chen J, Wang J, et al. Feature decomposition-based gaze estimation with auxiliary head pose regression[J]. *Pattern Recognition Letters*, 2024, 185: 137-142.
 - [11] Zhang X, Sugano Y, Fritz M, et al. It's written all over your face: Full-face appearance-based gaze estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, Hawaii, USA: IEEE, 2017: 51—60.
 - [12] Abdelrahman A A, Hempel T, Khalifa A, et al. L2cs-net: Fine-grained gaze estimation in unconstrained environments [C]//2023 8th International Conference on Frontiers of Signal Processing. Corfu, Greece: IEEE, 2023: 98—102.
 - [13] Chen Z, Shi B E. Appearance-based gaze estimation using dilated-convolutions[C]//14th Asian Conference on Computer Vision. Perth, Australia: Springer International Publishing, 2018: 309—324.
 - [14] Abdelrahman A A, Hempel T, Khalifa A, et al. Mobgazenet: robust gaze estimation mobile network based on progressive attention mechanisms[J]. *Machine Vision and Applications*, 2025, 36(4): 76.
 - [15] Huang G, Shi J, Xu J, et al. Gaze estimation by attention-induced hierarchical variational auto-encoder[J]. *IEEE Transactions on Cybernetics*, 2023, 54(4): 2592—2605.
 - [16] Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations[C]//37th International Conference on Machine Learning. Vienna, Austria: PMLR, 2020: 1597—1607.
 - [17] Grill J B, Strub F, Altché F, et al. Bootstrap your own latent: a new approach to self-supervised learning[J]. *Advances in Neural Information Processing Systems*, 2020, 33: 21271—21284.
 - [18] Mahmud Z, Hungler P, Etemad A. Gaze estimation with eye region segmentation and self-supervised multistream learning [J]. *arXiv preprint arXiv:2112.07878*, 2021.
 - [19] Du L, Zhang X, Lan G. Unsupervised gaze-aware contrastive learning with subject-specific condition[J]. *arXiv preprint arXiv: 2309.04506*, 2023.
 - [20] Cheng Y, Lu F. GazeTR: gaze estimation with transformer [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, Louisiana, USA: IEEE, 2022: 16347—16356.
 - [21] Hu M, Feng J, Hua J, et al. Online convolutional reparameterization[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, Louisiana, USA: IEEE, 2022: 568—577.
 - [22] Zhang X, Sugano Y, Fritz M, et al. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 41(1): 162—175.
 - [23] Funes Mora K A, Monay F, Odobez J M. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras[C]//Proceedings of the Symposium on Eye Tracking Research and Applications. Safety Harbor, Florida, USA: ACM, 2014: 255—258.
 - [24] Zhang X, Park S, Beeler T, et al. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation[C]//16th European Conference on Computer Vision. Glasgow, UK: Springer International Publishing, 2020: 365—381.
 - [25] Cheng Y, Wang H, Bao Y, et al. Appearance-based gaze estimation with deep learning: A review and benchmark[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, 46(12): 7509—7528.
 - [26] Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation[C]//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015: 1520—1528.
 - [27] Zhang C, Liu H, Deng Y, et al. Tokenhpe: Learning orientation tokens for efficient head pose estimation via transformers [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, BC, Canada: IEEE, 2023: 8897—8906.
 - [28] Kingma D P, Ba J. Adam: A method for stochastic optimization [J]. *arXiv preprint arXiv:1412.6980*, 2014.
 - [29] Zhang X, Sugano Y, Fritz M, et al. Appearance-based gaze estimation in the wild[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, Massachusetts, USA: IEEE, 2015: 4511—4520.
 - [30] Park S, Spurr A, Hilliges O. Deep pictorial gaze estimation [C]//15th European Conference on Computer Vision. Munich, Germany: Springer International Publishing, 2018: 721—738.
 - [31] Cheng Y, Lu F, Zhang X. Appearance-based gaze estimation via evaluation-guided asymmetric regression[C]//15th European Conference on Computer Vision. Munich, Germany: Springer International Publishing, 2018: 100—115.
 - [32] Y. Cheng, S. Huang, F. Wang, et al. A Coarse-to-fine adaptive network for appearance-based gaze estimation[C]//34th AAAI Conference on Artificial Intelligence. New York, NY, USA: AAAI, 2020: 10623—10630.
 - [33] Zhang X, Sugano Y, Bulling A, et al. Learning-based region selection for end-to-end gaze estimation[C]//31st British Machine Vision Conference. Virtual: British Machine Vision Association, 2020: 86.
 - [34] Van Der Maaten L, Hinton G. Visualizing data using t-sne[J]. *Journal of Machine Learning Research*, 2008, 9(11): 2579—2605.