Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Intelligent Manufacturing—Article

基于参考网格的差分眼部外观网络的视线估计

顾菸 [a], 王力翠 [b,*], 何龙 [a], 何先定 [a], 王建 [a]

[a] Chengdu Aeronautic Polytechnic, Chengdu 610100, China
[b] Department of Production Engineering, KTH Royal Institute of Technology, Stockholm 10044, Sweden

ARTICLE INFO

Article history:
Received 8 November 2019
Revised 11 June 2020
Accepted 6 August 2020
Available online 30 April 2021

关键词
视线估计
视线方向差异
孪生神经网络
跨样本估计
人机协作

摘要

人类的视线可以有效地传递人们的意图，因此，视线估计方法是智能制造中意图传递的重要研究内容。很多方法通过分析眼部图像，称为眼部图像片，实现视线方向的回归运算。但是，由于眼部图像存在个体差异，这类方法很难建立一个样本无关模型进行视线估计。在本文中，作者假设人眼的外观差异与视线方向差异有直接联系。基于这个假设，本文利用双眼眼部图像片在不同视线时的图像差异估计相应两种视线的差值，构建了差分眼部外观网络（differential eyes' appearances network, DEANet），并在公共数据集中进行训练。本文提出的DEANet主要基于孪生神经网络（Siamese neural network, SNNet）构建，包含两个结构相同的网络分支。多流数据分别输入到此孪生神经网络的两个分支中。两个网络分支共享相同的权值，实现眼部图像片的特征提取，然后对特征进行拼接，从而获得视线方向的差异。只要完成了视线方向差异模型的训练，在少量的校准图像片的情况下，就可以对其他样本的视线方向差异进行估计。由于测试阶段包含了被测试者的眼部信息，因此估计精度进一步提高。此外，本文提出的方法还有效地避免了在训练样本相关模型时需要大量数据的问题。本文还提出了一种参考网格策略，以便在测试阶段有效地选择一些参考眼部图像片，将它们作为网络的一部分输入，从而进一步提高估计精度。在公共数据集上的实验表明，本文提出的方法优于当前的方法。

1. 引言

在人类交流过程中眼神有着丰富的信息量。当在一个嘈杂的共同空间工作时，人们更喜欢通过眼神和手势等非语言行为来表达他们的意图。眼神中大量的信息有助于工作的完成。人的意图可以通过估计他（她）的视线方向有效地感知。许多学者开展了对于眼神估计的意图理解研究。例如，文献[1]中描述了如何通过眼神控制机器人将手中相应的物体递给人类的方法。这个实验表明，眼神中携带的丰富信息对协作有重要的影响。眼神估计已经被应用于许多领域，如人机协作（HRC）[1,2]、虚拟现实（VR）[3]和移动设备的控制器[4]。特别是在HRC中，除了手势、语音指令和身体运动[5,6]之外，眼神估计系统还可以通过多模态融合来控制机器人。眼神估计将扩大HRC的应用范围，有助于提高多模态机器人控制的可靠性。

在智能制造中，人类是具备智能和柔性特征的自动化[7,8]过程系统回路的组成部分，在与机器人的协作中发挥着重要作用。机器人可以处理的任务范围正在增加[9]，人类更倾向于通过自然方法与机器人交流。例如，

* Corresponding author.
E-mail address: lihuiw@kth.se (L. Wang).

英文原文: Engineering 2021, 7(6): 777–786
引用本文: Song Gu, Lihui Wang, Long He, Xianding He, Jian Wang. Gaze Estimation via a Differential Eyes' Appearances Network with a Reference Grid. *Engineering*, https://doi.org/10.1016/j.eng.2020.08.027

通过手势或眼神给机器人下达指令，而不是使用遥控器。此外，人们也不愿意使用侵入式的解决方案，比如戴着估计视线方向的特殊眼镜[10]。相反，可以在周围位置安装摄像头来观察操作员，通过分析摄像头中的数字图像估计操作员的视线方向。这是一种基于计算机视觉技术的常用的非侵入式方案。在对操作员的视线方向进行估计时，操作员不会感觉到设备的存在。

基于非侵入式的方案通常可以分为两种类型：基于模型的方法和基于外观的方法[11]。在基于模型的方法中，通过计算图像来估计眼睛各部分的几何模型，比如瞳孔的半径和瞳孔的中心位置，并基于几何模型[12,13]估计视线方向。在基于外观的方法中，通过分析眼部图像片直接回归出视线方向。一方面，与基于外观的方法相比，基于模型的方法，其视线方向估计的精度取决于采集图像的质量，如图像分辨率和亮度，用以确保精确地提取某些边缘或特征点。相比之下，基于外观的方法并不需要特征点。文献[14]对主流的视线估计方法进行了论证，证明了基于外观的方法比基于模型的方法更能获得较好的性能。另一方面，采用模型的方法需要先验知识，用来构建更加精确的视线估计模型[15]，这是一项具有挑战性的任务。此外，深度神经网络可以有效地获得数据的内在联系。深度神经网络在基于外观的方法中得到成功应用，大大提高了视线估计的精度。因此，近年来，基于外观的方法引起了广泛的关注[16–18]。文献[19,20]提出了基于视频的眼神估计系统，这是基于模型的方法。可以通过深度神经网络，如递归神经网络或长短时记忆神经网络，来增强系统的性能。然而，这类方法超出了本文的研究范围。

使用基于外观的方法，关键步骤是确定输入图像和视线方向之间的关系。很多学者建立了不同的模型来确定这种关系。这些模型利用从不同人员采集的数据样本进行训练和测试，即跨样本估计方法。相应的模型也被称为人员无关模型。因为人员无关模型并不包含被测试对象的视线信息，因此样本个体在外观上的差异将会影响估计的准确性。如果在模型的训练过程中，引入了测试过程中的某些条件，如被测试人员的外观、测试现场光照水平等，则模型系统的性能将得到提高。一种常用的方法是收集被测试人员的标记数据以进行模型训练。这被称为人员相关模型。然而，学习一个人员相关模型需要大量的标记数据，它是一项耗时的任务，限制了这种方法的适用性。虽然有些技术，如文献[21,22]中提出了数据收集复杂性的下降方法，但仍然需要大量的训练数据。受文献[23–25]的启发，本文提出采用差分图像和视线差异输出分别代替输入图像和输出的视线方向。一旦确立了两个输入图像的差异和两个视线方向差异之间的关系，就只需要新样本的少数标记图像，并且这些标记图像可以作为测试阶段的输入之一。通过这种方法，对视线方向进行精确的估计。

本文提出了一个基于深度神经网络学习框架的差分眼部外观网络来估计眼神方向。该网络基于孪生神经网络（SNNet）[26]，包含两个相同的分支。两个样本集作为两个分支输入到网络中。每个样本集包括一个人脸图像中左眼和右眼的图像片。作为多流结构[27]的一部分，这两个图像片分别输入到孪生网络的一个分支中。每个分支网络包含具有不同参数的VGG16网络[28]，从所有图像片中提取特征。将这两个分支的输出与头部位置信息拼接在一起。网络的输出是两个样本集的视线差异，之后进入全连接网络。在测试阶段，将被测试人员已标记的样本集作为参考样本集，输入到孪生网络的一个分支中。需要估计的样本集输入网络的另一个分支，网络的输出是参考样本集与被估计样本集之间的视线差。由于已经对参考样本集的视线方向进行标记，因此估计的视线方向等于网络输出的视线方向与参考样本集标记的视线方向之和。此外，可以利用参考样本集选择策略来进一步提高系统的性能。我们提出的方法假设人眼的外观差异与相应视线方向的差异有关。由于训练模型在测试阶段中嵌入了被测试人员的信息，因此提高了估计的精度。此外，在视线估计时，只需要少量的被测试者的标记图像。网络不需要大量数据来进行人员相关模型的训练。在当前主流数据集上的实验表明，本文提出的算法的性能优于当前其他方法。

本文的方法主要有以下的贡献：

（1）提出了一个眼部图像与归一化头部姿态信息相结合的新方法。多流信息输入孪生网络的不同分支中。孪生网络模型框架既包含测试阶段被测试人员的信息，又不需要收集大量数据来训练样本相关模型。

（2）提出了一种新的参考样本集选择策略，提高估计精度。在视线空间中构造了参考网格，并通过估计值直接选择有效的参考样本集，简化了系统的计算。

本文的内容组织：第2节介绍了相关的研究工作；第3节详细说明了本文提出的方法；第4节是实验结果和讨论内容；最后，第5节是论文结论和未来研究计划。

2. 相关研究工作

本节简要介绍基于外观的眼神估计、样本相关估计和孪生神经网络的最新进展。

2.1. 基于外观的眼神估计

大多数基于外观的视线估计算法主要是回归方法。估计的视线方向是输入图像的函数。直观上，眼部图像片含有丰富的视线方向（左眼和右眼）的信息，可用于估计视线方向。文献[29]提出了一种基于多模态卷积神经网络（CNN）的自然场景中眼部外观的视线估计方法。文献[30]中，Lian等提出了一个共享的CNN，估计从不同相机拍摄到的多视角眼部图像片的视线方向。Liu等在文献[23,25]中阐述了在差分CNN网络上直接训练，估计一对眼部图像片之间的视线差异的方法。Park等[31]提出了一种新的全卷积图像表示框架来估计视线方向。然而，除了眼部图像片外，许多其他因素也会影响估计的精度，如头部位置、图像中眼睛的比例、头部姿势等。Liu等[32]使用眼部图像片和眼部网格构建了一个两步训练网络，提高了在移动设备上的估计精度。Kyle等[4]将眼部图像片、面部图像片和面部网格作为系统的输入，获得了良好的性能。Wong等[33]构建了一个ResNet模型，该模型结合头部姿势和面部网格特征在移动设备上估计视线方向。文献[34]根据瞳孔中心的位置将视线分为三个区域，采用无监督学习方式构建网络来估计视线方向。Yu等[17]引入一个约束眼部基准模型，通过融合眼部基准位置实现视线估计。Funes-Mora和Odobez [35]提出了一种基于RGB-D相机的头部姿态不变性的视线估计算法，并在低分辨率数据集[36]对其性能进行了评估。Zhang等[16]在自己模型的基础上对上述所有的影响因素进行了分析。文献[37]采用全人脸图像作为系统的输入，并采用了具有空间权重的Alex-Net [38]网络，其效果明显优于许多采用眼部图像作为输入的算法。这些研究表明，全脸的外观比只有眼部图像外观的方法，在头部姿势和光照改变时稳定性更好。但是，全脸外观方法的输入数据比眼部图像外观的数据多得多，大大增加了计算的复杂度。文献[39]提出了在保持估计精度的同时有效地压缩图像数据量的方法。全脸的方法和眼部图像的方法哪个性能更好，目前尚未有明确定论。

不进行任何预处理就将原始图像输入系统会增加网络回归的复杂度。在预处理阶段对一些信息进行标准化可以降低网络复杂度。Sugano [40]提出一种新的归一化方法，在输入网络之前对图像进行对齐。其他数据，包括图像和视线方向，也被转换到归一化的空间。在网络训练或测试时，不需要考虑目标的尺度问题。文献[40]将相机从人眼转换到固定位置，构建虚拟相机，在虚拟相机坐标下获得视线方向。Zhang [41]详细分析了归一化方法，将归一化方法推广到文献[37]中的全脸图像。

2.2. 样本相关估计

很多视线估计算法的目标是训练一个样本无关的模型，达到良好的跨样本估计性能。在输入图像和视线方向之间建立一个样本无关模型描述这两者之间的关系。但是，文献[25]中提出不同样本眼部的视觉轴和光轴之间的关系是不同的。样本无关的模型不能准确描述这种视觉轴和光轴之间的关系，而样本相关模型可以准确估计视线方向。文献[16]证明了只要有足够的训练样本，就可以保证样本相关模型的性能。

样本搜集阶段相当耗时。近期的论文中，学者们提出了很多简化样本收集的方法。Sugano [42]提出了一种连续更新估计参数的增量学习方法。文献[43]中，不同设备端收集到的数据被输入一个CNN网络中，该网络包含共享的特征提取网络层，以及设备特定的编码器/解码器。Huang [22]建立了一个监督自学习算法来逐步地训练视线模型。并且，数据验证的鲁棒机制可以区分良好的训练数据和噪声数据。Lu [21]等提出了一种自适应线性回归方法，自适应地选择一组最优样本进行训练。在所需的训练样本的数量显著减少的同时，仍然保持了较好的估计精度。虽然上述方法简化了数据收集过程，但许多方法仍然需要标记样本来训练特定的模型。Yu [44]基于少量样本生成大量的标记数据，设计了一个视线方向二次计算框架。Liu等[23]提出了一种仅基于一个眼部图像片的特定样本视线估计的新方法。根据输入的图像，利用SNNet来估计视线方向的差异。SNNet网络训练之后，在测试阶段需要引入一定的标记样本。

2.3. 孪生神经网络

SNNet首次在文献[26]中用于验证平板电脑上的手写输入签名。SNNet的特征之一是它包含两个相同的分支。相对于单个输入，SNNet网络的输入是一对具有相同类型和不同参数的输入。因此，网络的输出是相应输入的差异。该方法被用在许多领域。Venturelli [24]提出利用SNNet框架在训练阶段估计头部姿态。为了提高回

归网络的学习能力，在损失函数中增加了一个差异学习项。Varga等[45]采用孪生网络架构，减小在三维人体姿态估计中对数据增强的需求。文献[23,25]提出的方法与本文最相似。然而，在这些参考文献中没有讨论双眼和头部姿态对网络的影响。同时，文献的两种算法都证明了参考样本对估计精度的影响。但是，在参考文献[23,25]中并没有系统地讨论参考样本的选择策略。由于孪生网络的输入需要一对数据集，其训练样本的组合可能性使得训练数量极速增加。在文献[46–48]中分析了训练样本中训练子集的选择策略。

3. 差分眼部外观网络

虽然本文提出的是样本无关模型，但在测试阶段会同时考虑测试人员的样本信息。系统的框架如图1所示。整体上，系统的整个框架是基于一个SNNet构建的。系统没有采用单一输入，而将是一个信息对，分别输入到网络中的两个分支。并且，这两个分支共享相同的权重。待估计的人脸图像和参考人脸图像作为系统的原始输入。每个图像通过原始头部姿态信息$\tilde{H}$，可以被归一化为左眼图像片和右眼图像片。所有的归一化图像片都包含在孪生网络的的输入对中，分别是参考样本集P_f和估计样本集P_t。每个样本集包括一个左眼眼部图像片I^l、一个右眼眼部图像片I^r和归一化的头部姿态信息H。参考样本集对应的视线方向被提前标记，称为参考视线$\tilde{G}_f$。系统的输出$\tilde{G}_t$，称为估计视线，是与测试样本集对应的视线方向。所有的图像和$\tilde{G}_f$，都通过原始的头部姿态信息进行归一化。在归一化过程中，被估计人脸图像和参考人脸图像将使用不同的原始头部姿态信息，在图1中分别以$N(\tilde{H}_t)$和$N(\tilde{H}_f)$表示。所有通过归一化校准的图像片都被送入DEANet。待测试的归一化视线就是网络的输出和归一化参考视线的和，然后进行去归一化$N^{-1}(\tilde{H}_t)$，去归一化是在相同参数下的归一化$N(\tilde{H}_t)$的反操作。

3.1. 定义

对视线方向表示方法可以分为两类：二维和三维表示。二维的视线位置由屏幕上视线位置的坐标来表示，多用于移动设备的显示装置。三维视线方向是在三维空间中从参考点到目标点的方向。它由相机坐标系中的三个角度组成：偏航、俯仰和滚转。实际操作中，三维视线方向被定义为从参考点到目标点的单位向量。本文通过球坐标系进行简化，包括φ和θ，$G=[\varphi^g, \theta^g]'$。同时，参考点设为眼睛的中心。具体来说，本文主要是对三维视线方向进行估计，将三维视线方向定义为从左眼中心到目标点的向量。三维视线与二维视线可以互相转化。如果能在三维空间中得到二位的平面，就可以从三维视线方向得到二维视线的位置。同样的，对头部的姿态定义采用三维视线方向相同的方法，即$H=[\varphi^h, \theta^h]'$。

3.2. 预处理和归一化

采用文献[37,40]中的方法，对原始图像归一化从而进行视线估计，可以减轻摄像机不同和原始头部姿态信息的影响，从而降低网络复杂性。归一化过程是一系列的透视转换过程，以获得归一化图像片与从统一参考点的虚拟相机中拍摄的图像的一致性。文献[40,41]对归一

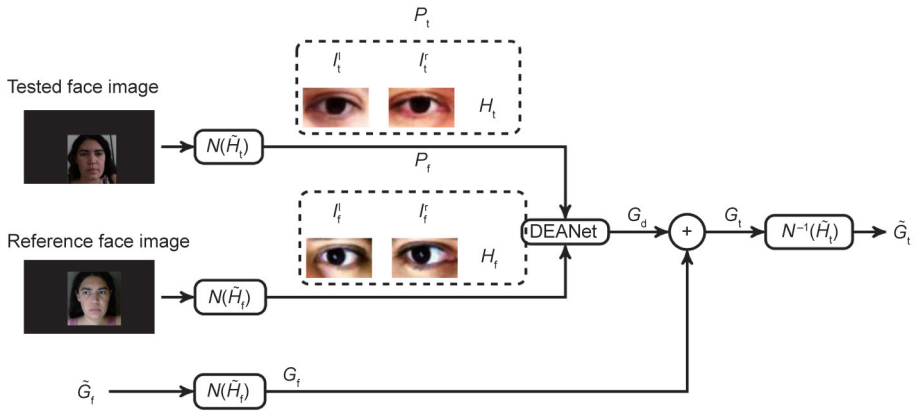

图1. 文中提出的框架结构。被测试的人脸图像和参考的人脸图像分别根据原始的头部姿态信息进行归一化，构建孪生图像对P_t和P_f。每对孪生图像包括一个左眼眼部图像片I^l、一个右眼眼部图像片I^r和归一化的头部姿态信息H，其中，$P_t=\{I_t^l, I_t^r, H_t\}$，$P_f=\{I_f^l, I_f^r, H_f\}$。对原始的参考视线$\tilde{G}_f$进行标记，归一化得到$G_f$。归一化的图像数据输入DEANet神经网络，回归出P_t和P_f的视线差异G_d。$N(\tilde{H}_t)$和$N(\tilde{H}_f)$是相同的归一化操作，只是参数不同。$N^{-1}(\tilde{H}_t)$是去归一化操作，去归一化是在相同参数下的归一化$N(\tilde{H}_t)$的反操作。

化步骤和性能做出了详细的说明。本节对关键步骤进行介绍。

首先，对单幅脸部图像，如图1中的测试图像进行处理。采用主流算法[49]对脸部关键点，如眼睛和嘴巴的角点，进行检测。由角点计算的左眼中心点、右眼中心点、嘴部中心点构造平面。从右眼中心到左眼中心的连线是x轴，y轴在平面内垂直于x轴，从眼睛指向嘴。z轴在平面内按照右手规则获得。三轴与左眼中心或右眼中心合并为坐标原点，构成双眼的归一化空间。根据检测到的面部关键点和一般面部形状模型[16]，通过EPnP算法[50]计算归一化的头部姿态信息。注意主流数据集中提供了原始的头部姿态信息和相机的内参，其性能将在第4节中进行评估。所有输入到网络的图像片都必须经过归一化，投影到归一化空间中。为了减少图像光照对系统的影响，对所有归一化后的图像片进行直方图均衡。

DEANet对于归一化有两个优点：

（1）归一化作为一种图像对齐的操作，降低了网络的复杂性，减少了不同相机成像距离、相机内参和原始头部姿态信息对眼部图像片的影响。归一化的图像可以同时输入到孪生网络中，孪生网络中的分支共享相同的权重。

（2）归一化简化了视线差异的计算。无论坐标如何变换，所有的参数都在归一化空间中，并且视线差的计算等价于对两个视线向量的操作。参考样本的选择策略将在3.4节介绍。

3.3. 网络的训练

无论照相机的内参和图像大小如何变化，归一化后，所有图像片都将在归一化空间中对齐。归一化图像片作为网络输入使网络学习更有效率，以提高系统的性能。我们假设，人每个眼睛的外观差异与相应注视方向的差异有关。并且，这种关系通用于所有人。为此，本文提出了一种基于SNNet的外观视线估计方法。网络的结构和配置如图2所示。

在训练过程中，采用一对样本集P_t和P_f作为DEANet神经网络的输入。每一对样本都包括左眼图像片、右眼图像片和归一化的头部姿态信息。两个样本集分别输入到SNNet的两个共享参数的分支中。在孪生网络的分支中，所有输入的图像片均固定大小的36×60 RGB或灰度图像。若输入为灰度图像，将把它作为三个通道中具有相同强度值的RGB图像处理。归一化的头部姿态信息是一个长度为2的向量。左眼图像片和右眼图像片分别输入VGG16网络，提取两个图像片的特征，得到长度为512的向量。VGG16网络之后接着一个系列操作，包括大小为1024的全连接层（FC）、批处理归一化（BN）和ReLU激活。将通过图像片计算得到的特征图拼接起来，并连接一个尺寸为512的FC层。将归一化的头部姿态信息加入此特征向量中，并通过BN、ReLU激活、256的FC层和另一个ReLU激活。最后，将两个孪生分支计算出的特征图拼接起来，依次送入尺寸为256和2的FC层。为了避免过拟合，在最后一个FC层之前添加了一个dropout层。

3.3.1. 用于训练的样本集的选择

根据本文的假设，将属于同一个人的一对已标记的样本集输入网络。如果数据集有N个训练样本，就会有N^2种可用于网络训练的样本对的组合。与单输入算法[4,37]相比，由于网络框架的不同，本文提出的方法会有大量的训练样本。由于样本量足够，因此在训练阶段采用了训练样本的一个子集。文献[47,48]讨论了子集的选择方法。但是由于文献讨论的是分类任务，因此其中的子集都是由正样本对和负样本对组成。然而，本文提出的是一种回归方法，不能使用具体的正样本对和负样本对。本文的方法中，训练过程采用随机选择$K < N^2$对训练样本。

3.3.2. 损耗函数

根据图1，当给定G_f时，如果DEANet估计的视线差异接近标记的视线差异，估计的G_t将接近G_t^{gt}。如果有K对标记训练样本$\{P_{t,k}, G_{t,k}^{gt}\}_1^K$和$\{P_{f,k}, G_{f,k}^{gt}\}_1^K$，其中，$G_{t,k}^{gt} \in \mathbb{R}^{2\times1}$和$G_{f,k}^{gt} \in \mathbb{R}^{2\times1}$分别对应于$P_{t,k}$和$P_{f,k}$的标记数据。损失函数构造如下：

$$L = \frac{1}{K}\sum_{k=0}^{K} ||G_{d,k} - G_{d,k}^{gt}||_2^2 \qquad (1)$$

式中，标记的视线差异$G_{d,k}^{gt} = G_{t,k}^{gt} - G_{f,k}^{gt}$和$G_{d,k}$是神经网络基于$P_{t,k}$和$P_{f,k}$预测的视线差异；$\|\cdot\|_2$是$l_2$-范数的操作符号。

3.4. 测试阶段的参考网格

图1显示了在测试阶段如何通过一个已经标记的参考样本集进行视线估计的过程。参考样本集的选择将影

Input	P_t			P_f		
	H_t 2 × 1	I_t^l 3@36 × 60	I_t^r 3@36 × 60	I_f^l 3@36 × 60	I_f^r 3@36 × 60	H_f 2 × 1
DEANet		VGG16 FC-1024 BN-1024 ReLU	VGG16 FC-1024 BN-1024 ReLU	VGG16 FC-1024 BN-1024 ReLU	VGG16 FC-1024 BN-1024 ReLU	
		CAT FC-512		CAT FC-512		
	CAT BN-514 ReLU FC-256 ReLU			CAT BN-514 ReLU FC-256 ReLU		
	CAT					
	FC-256					
	ReLU					
	Dropout-0.5					
	FC-2					
Output	G_d 2 × 1					

图2. DEANet网络结构（从上到下）。I_t^l、I_t^r、I_f^l和I_f^r是大小为36×60的RGB图像。H_t和H_f是与双眼图像片对应的归一化的头部姿态信息。G_d是估计视线差异。它们都是长度为2的向量。FC是全连接层，BN是批处理归一化层，ReLU是ReLU激活，Dropout是Dropout层。每层网络通过它们的参数命名。CAT是将两个向量拼接到一个向量中的操作。共享相同权重的层以相同的颜色突出显示。

响视线估计的精度。直观上，一个优良的参考样本选择方法选择的参考样本集与被测试的样本集之间的差异不应太大。较大的差异将导致在估计过程中产生很大的误差。此外，在测试阶段，采用多个参考样本集进行估计，估计精度会优于仅用单个参考样本集的结果。这个结论将在4.3节进行说明。根据上述结论，在整个视线空间中构造一个参考网格，包括视线方向的两个维度分量，如图3所示。当输入图像片之间的差异很小时，DEANet网络的输出也很小，反之亦然。因此，DEANet网络的输出（即视线差异）是参考图像片和被测试图像片之间距离的度量。如果参考数据满足均匀分布，如图3所示，并且网格的步长足够小，那么参考样本图像片和测试图像片之间的差异也将足够小，足以获得良好的精度。例如，12个红点是参考视线的候选点，表示为$G_{f,j}, j = 0, 1, \ldots, 11$。估计视线为一个蓝色点，表示为$G_t$。显然，相较于其他参考点，$G_t$是通过$G_{f,3}$、$G_{f,4}$、$G_{f,6}$和$G_{f,7}$来计算的，因为$G_t$和这四个参考视线点中任何一个之间的距离小于和其他参考视线点之间的距离。同时，由于估计视线与参考视线在视线空间中的距离可以通过本文提出的DEANet网络中的视线差值来进行预测，因此采用视线差值小于一定阈值的参考视线，来预测估计视线。为了避免经验参数，本文采用了4个参考视线点，这4个

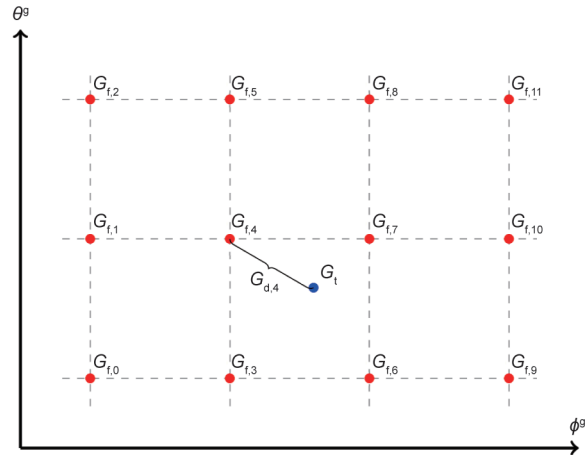

图3. 视线空间中的参考网格的示例。视线空间中分布着12个参考视线（标记在红色点上）。蓝色点代表待估计视线。视线空间中，$G_{f,j}$和G_t之间的距离是由相应的视线差异$G_{d,j}$来预测的。

参考点对应的视线差异小于其他参考点。之后，通过将每个参考视线添加到相应的视线差异中来预测估计视线。最终的估计值就是它们的平均值。实验证明，该方法对于所有测试集都取得了较好的效果。

文献[25]中，平均权重是通过比较输入图像片中提取的两个特征图来确定的。根据DEANet的结构，网络的输出与两个图像片的差异有关。相对于文献[25]中提出的特征图，使用视线差异作为参考选择的标准简化了计算。

4. 实验

4.1. 算法实现细节

本文提出的DEANet网络是在pytorch平台下搭建的。针对每个测试者随机选择10 000对训练样本进行训练。采用迁移学习，通过预训练模型[28]初始化VGG16模型的参数。采用了动量为0.9的随机梯度下降（SGD）优化方法，权重衰减为0.0001。训练样本批次为512。初始学习率为0.1，每训练5次学习率衰减0.1。网络使用了1个GTX 1080 ti的GPU，每个人训练20次。

本节包含三个实验：第一个实验（见4.3节）基于MPIIGaze数据集评估验证了DEANet网络，并阐述参考集的选择策略；第二个实验（见4.4节）评估了DEANet网络在跨人群样本和跨数据集的预测表现；第三个实验（见4.5节）评估了DEANet的抗噪性能。

4.2. 数据集和评估标准

在两个公共数据集MPIIGaze和UT-Multiview上评

估了DEANet网络的性能。MPIIGaze首次出现在文献[16]中。它包括来自15名不同年龄和性别的参与者的213 659张图片。这些图像在不同时间段内收集。为了评估本文提出的DEANet网络在RGB图像中的性能，我们仍对MPIIGaze数据集中的眼部图像片和视线方向进行了归一化。同时，在归一化过程中，直接使用了数据集提供的原始头部姿态信息和目标信息。文献[40]中首次提出了UT-Multiview。它包括来自50个人的64 000张原始图像。这个数据集通过三维眼睛形状模型生成了大量的眼睛图像样本。UT-Multiview比MPIIGaze具有更宽的视线角度分布范围。由于本文主要基于文献[40]的方法对图像进行归一化处理，所以归一化图像片的大小与UT-Multiview中的相同。将UT-Multiview中的所有灰度图像片作为DEANet网络的训练样本，用来评估网络的性能。

实验中，MPIIGaze数据集使用了留一法（leave-one-person-out）标准，UT-Multiview数据集使用了三折叠交叉验证法（three-fold cross-person）评估标准。本节中采用的标准与其他最先进的算法相同[4,16,18,25,37,40]。

4.3. 参考样本集的选取

根据上述描述，参考样本集的性能将影响系统的估计精度，样本集是DEANet网络的一个关键因素。在本实验中，在MPIIGaze数据集中每个人随机选取500个参考样本作为参考样本集。参考样本集和属于同一个人的图像样本组成了要进行测试的双眼图像片并将其输入到网络中。图4显示了每个人的平均角度差异。将每个人的所有双眼图像片输入DEANet网络进行视线估计，每个参考样本的平均角度差异计算公式如下：

$$A_{\mathrm{t}}=\frac{1}{M}\sum_{m=0}^{M}\omega\left(G_{\mathrm{t},m},G_{\mathrm{t},m}^{\mathrm{gt}}\right) \qquad (2)$$

式中，M是数据集中每个人的样本数；$\omega(\cdot,\cdot)$是计算两个向量之间角度差异的函数。公式（2）中ω函数是估计差异的另一个度量，等价于公式(1)中的l_2范数。如图4中的蓝色条所示，每个人都有不同的估计精度。有些人，如0、1、2号，他们的平均角度差异比其他人要小。同时，其他人的平均角度误差，如3、7、8、9号，都比之前那些人的情况要差得多。例如，一些人（7号）的眼部图像中有眼镜，而其他人则没有。如果选用的参考样本集中没有眼镜信息，而测试样本集包含了眼镜信息，由于眼镜信息作为噪声进入到了估计中，这样不同外观将导致估计精度的严重下降。这一点在图5（d）和（e）中

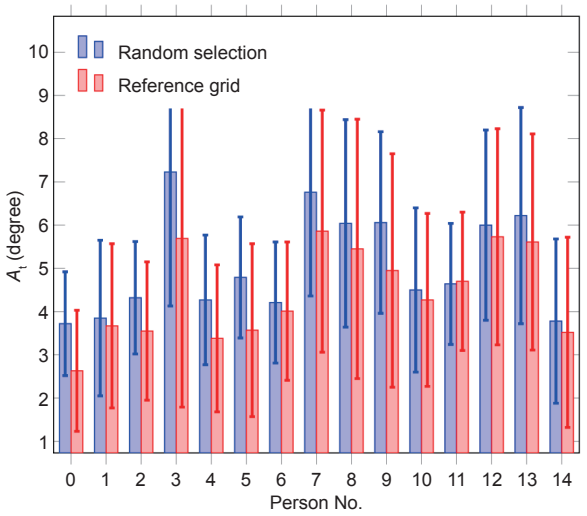

图4. 不同参考选择策略下MPIIGaze数据集中每个参考的平均角度差异：随机选择策略，随机选用500个参考样本集；参考网格策略，由参考网格确定选用的12个参考样本。

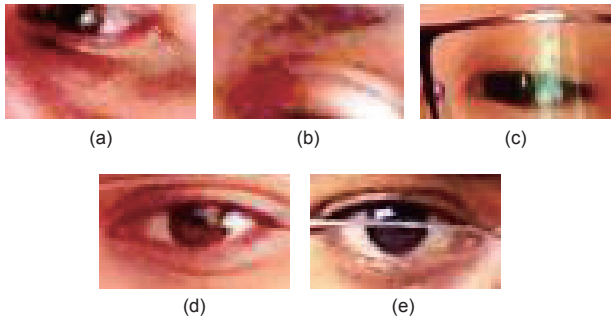

图5. 导致较大估计误差的归一化图像片。（a）、（b）不准确的归一化眼部图像片（p03-day54-0097-left and p08-day31-0301-left）；（c）眼镜引起的噪声（p09-day12-0158-left）；（d）、（e）无眼镜参考样本集的图像（p07-day24-0046-left）和眼镜测试样本集的图像（p07-day25-0255-right）。每个眼部图像片的名称都来自于MPIIGaze数据集。

体现。图5（a）和（b）作为另一个例子说明了归一化对系统的影响。虽然文献[16]证明了在归一化阶段，使用一般平均面部形状模型能够准确地估计视线方向，但是如果归一化后眼部图像效果不佳，在估计阶段将产生很大的误差。具体实例如图5所示。

良好的参考样本集的选择策略有助于系统性能的提高。参考样本集选择策略的一个关键因素是确定哪些图像片是参考样本集的候选图像片，哪些不是。这与被估计样品的分布有关。图6表示了0、5、7号中，在视线空间中随机选择的500个参考样本集的分布。每个参考视线都可以用视线空间中的一个点来表示。对于参考i，当平均角度差异$A_{\mathrm{t},i}$小于所有参考的平均值时，相应参考被认为是“好”参考（在图6中以红色标记）。相反，当$A_{\mathrm{t},i}$大于所有参考的平均值时，相应参考被认为是“坏”

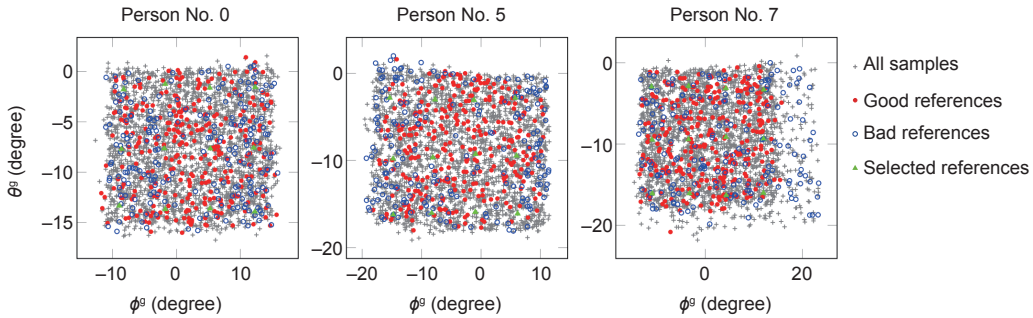

图6. MPIIGaze数据集中0、5、7号视线角度分布。任何参考样本集都可以用视线角度空间中标注的一个点来表示其视线方向。红色点是“好”的参考样本集，其值A_t小于所有参考的平均值；蓝色点是指值大于所有参考平均值的“坏”参考样本集。灰色点是每个人的所有样本。绿色点是实验中参考网格确定的参考点。

参考（在图6中以蓝色标记）。灰色点代表每个人整体分布的所有样本。图6中可以看出，“坏”参考（如7号）几乎都位于整个分布的外围，而“好”参考则均匀地分布在整个空间中。一些包含较大视线方向的样本集，不能作为参考样本集。此外，单一的参考选择策略并不足以提供准确的估计。

图6表明，参考样本集的分布会影响系统的性能。此外，参考样本集和被估计样本集之间的差异也对系统的性能有影响。值得注意的是，两个样本集之间的真实差异可以用本文提出网络的G_t^{gt}和G_f^{gt}之间的实际差异来表示。此外，该系统的估计差异还可以表示为$\omega(G_t, G_f^{gt})$。这也是两个样本集之间的真实差异的预测值。图7表示了两个样本集的差异与估计精度之间的关系。为了对图示进行简化，$\omega(G_t^{gt}, G_f^{gt})$被量化为100个bin，$\omega(G_t, G_f^{gt})$为相应的平均值，分别在图7中表示为$\bar{\omega}(G_t^{gt}, G_f^{gt})$和$\bar{\omega}(G_t, G_f^{gt})$。当被估计视线与参考视线之间的差异增加时，估计误差就会增大。接近被估计视线方向的“好”的参考视线方向将获得较好的估计精度。由于没有提前获得估计样本的视线方向，因此将会需要更多数量的参考样本集。这涉及在参考样本的数量与估计精度之间的权衡。此外，虽然没有提前获得估计视线方向，但可以预测获得估计视线方向的取值范围。根据视线方向的取值范围就可以构建参考网格。为了能在所有实验中都取得良好的性能，本文建立了一个三行四列的参考网格。如图6中绿色的圆点所示。在此基础上，使用MPIIGaze数据集，对每个人在具有参考网格的DEANet网络上进行了评估；平均角度差异如图4所示（红色条）。结果表明，几乎所有具有参考网格策略的平均角度差异都优于随机选择策略的差异。对所有人的平均角度差异从随机选择策略的5.09下降到参考网格策略的4.38，因此使用参考网格策略使性能提高了14%。

4.4. 跨人群样本和跨数据集的评估

DEANet网络是一个样本无关模型，可以估计不同人的视线方向。被测试样本的信息在测试阶段作为参考样本集输入网络。因此，有效地避免了训练样本与测试样本不同的问题。为了评估DEANet网络在这一问题上的性能，本文在两个公共数据集上都进行了跨样本评估。表1表示了本文的算法和其他算法在MPIIGaze和UT-Multiview数据集上的平均角度差异。本文提出的算法在这两个数据集中都取得了较好的效果。尽管文献[25]和本文算法都采用了结构相同的SNNet网络，但本文算法的性能优于文献[25]，本文算法涉及更多的信息，包括眼睛和头部姿态的信息。与MPIIGaze相比，UT-Multiview数据集包含了更多的人，因此在UT500 Multiview上评估的所有算法的性能都优于在MPIIGaze上评估的算法。作为数据驱动模型，训练数据的多样性增加了预训练模型的性能，本文提出的DEANet网络在这两个数据集上的性能都优于其他算法。

为了证明所提出方法的鲁棒性，本文进行了跨数据集的评估。模型在UT-Multiview数据集上进行训练，然后在MPIIGaze数据集上进行测试。图8表示了所有跨数据集评估算法的平均角度差异[16,29,40,51,52]。由于训练样本的视线分布不同于测试样本的分布，所有算法在跨数据集评估中的性能都弱于跨样本评估的性能。然而，我们提出的DEANet是一个差分网络，网络的输入和输出均是差分的输入与输出。本文的方法比其他传统方法更具有对视线分布的鲁棒性。我们提出的方法的平均角度差异为7.77°，标准差为3.5°。

4.5. 网络的抗噪性能

在之前的评估中，我们提出的DEANet网络在视线估计方面取得了很好的表现。本节进一步研究网络在噪

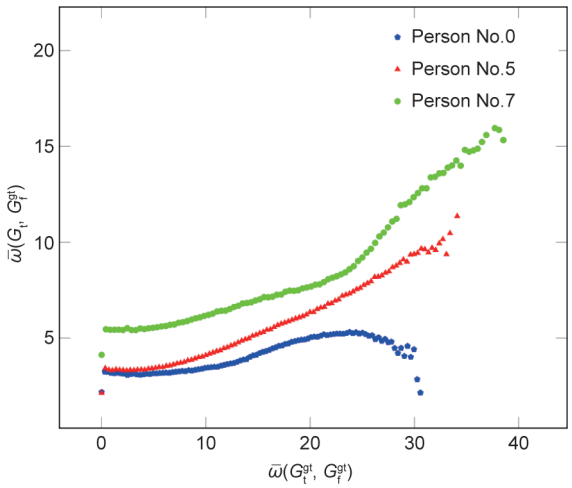

图7. 0、5、7号估计差异（y轴）与两个样本集（x轴）之间的关系。

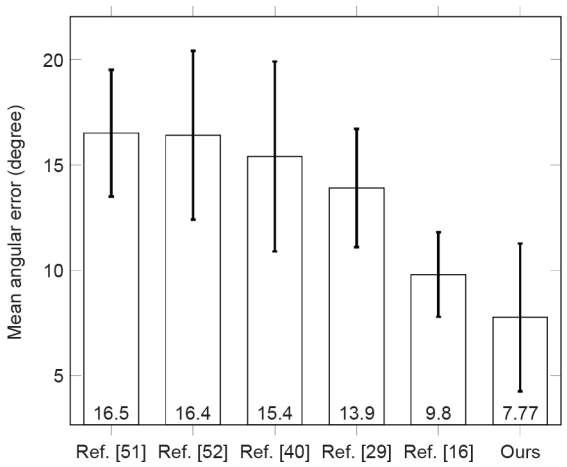

图8. 对UT-Multiview数据集进行训练和对MPIIGaze数据集进行测试的跨数据集评估的平均角度差异。

表1 两个具有平均角度差异（度）的常用数据集上的视线方向结果

Method	MPIIGaze	UT-Multiview
GazeNet [16]	5.5	4.4
Diff-NN [25]	4.64	4.13
RT-GENE Net [18]	4.8	–
iTracker [4]	5.6	–
Full face [37]	4.8	–
MnistNet [29]	6.1	–
LbS [40]	6.7	6.5
Ours	4.38	3.56

A leave-one-person-out protocol was used in the MPIIGaze dataset, and a three-fold cross-person validation protocol was used in the UT-Multiview dataset.

表2 图像分辨率的影响。在具有不同图像分辨率的MPIIGaze和UT-Multiview数据集上评估了平均角度差异

image resolution	MPIIGaze		UT-Multiview	
	Ours	GazeNet [16]	Ours	GazeNet [16]
18×30	5.41	–	3.75	9.9
9×15	8.57	–	5.42	11.4
5×8	12.10	–	13.07	15.7
Average	8.69	11.7	7.41	12.3

声情况下的性能，如头部姿态信息的影响和图像分辨率的影响。为了处理DEANet网络中任意的头部姿态信息，这里采用了归一化的头部姿态信息。为了证明人体头部姿态信息对DEANet的影响，我们在没有头部姿态信息的MPIIGaze数据集中进行了跨样本评估。本实验在MPIIGaze数据集上重新训练了一种没有头部姿态信息的新网络。如表2所示，评估的所有样本的平均角度差异为4.46°，略高于具有头部姿态信息的网络（4.38°）。如果没有头部姿态信息，网络的性能将会略有下降。对于DEANet等深度网络来说，头部姿态信息作用甚微。然而，对于一个较浅的网络，这些信息仍然很重要，如在文献[16]中对MnistNet [53]进行的评估。浅层网络通常应用在远程设备中，以节省计算资源。

此外，实验还研究了图像分辨率对视线估计的影响。实验采用了与第4.4节所述参数相同的网络参数，并进行了跨样本评估。也采用了与4.4节相同的评估标准。在视线估计中，所有图像片的大小设置为18×30、9×15和5×8。同时，通过插值将不同大小的图像片恢复到原来图像片的大小（36×60），以适应网络的输入。如表2所示，在不同图像分辨率下将DEANet和GazeNet [16]网络的性能，在MPIIGaze和UT-Multiview数据集上进行了比较。实验显示，DEANet的性能优于GazeNet。

5. 结论

本文提出了一种基于外观的视线估计新方法。三个数据流，包括双眼的眼部图像片和头部姿态信息同时输入神经网络，并基于SNNet网络框架训练了一个样本无关的模型。由于采用了视线差异的方法，因此可以在测试阶段使用被测试者的特定信息。同时，为参考点建立了参考网格，采用参考点选择策略提高了系统估计精度。本文的方法在两个公共数据集上进行了评估：MPIIGaze和UT-Multiview。实验结果表明，本文的方法比其他方法取得了更优异的性能。

所有的实验都在公共数据集上进行了理论分析。本文提出的方法将作为多模态融合的人机协作中的一种模态应用于人机交互控制中，这也是我们下一步的研究方向。

致谢

本文得到了四川省科技计划项目（2018SZ0357）的支持和国家留学基金管理委员会的资助。

Compliance with ethics guidelines

Song Gu, Lihui Wang, Long He, Xianding He, and Jian Wang declare that they have no conflict of interest or financial conflicts to disclose.

References

[1] Palinko O, Rea F, Sandini G, Sciutti A. Robot reading human gaze: why eye tracking is better than head tracking for human-robot collaboration. In: Proceedings of 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2016 Oct 9–14; Daejeon, Republic of Korea. New York: IEEE; 2016. p. 5048–54.
[2] Duarte NF, Rakovic M, Tasevski J, Coco MI, Billard A, Santos-Victor J. Action anticipation: reading the intentions of humans and robots. IEEE Robot Autom Lett 2018;3(4):4132–9.
[3] Thies J, Zollhöfer M, Stamminger M, Theobalt C, Nießner M. FaceVR: real-time facial reenactment and eye gaze control in virtual reality. ACM T Graphic 2018;37(2):1–15.
[4] Krafka K, Khosla A, Kellnhofer P, Kannan H, Bhandarkar S, Matusik W, et al. Eye tracking for everyone. In: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30; Las Vegas, NV, USA. New York: IEEE; 2016. p. 2176–84.
[5] Liu H, Wang L. Gesture recognition for human-robot collaboration: a review. Int J Ind Ergon 2018;68:355–67.
[6] Liu H, Wang L. Human motion prediction for human-robot collaboration. J Manuf Syst 2017;44(Pt 2):287–94.
[7] Wang L. From intelligence science to intelligent manufacturing. Engineering 2019;5(4):615–8.
[8] Liu H, Fang T, Zhou T, Wang L. Towards robust human–robot collaborative manufacturing: multimodal fusion. IEEE Access 2018;6:74762–71.
[9] Day CP. Robotics in industry-their role in intelligent manufacturing. Engineering 2018;4(4):440–5.
[10] Bulling A, Roggen D, Tröster G, Tröster G. Wearable EOG goggles: seamless sensing and context-awareness in everyday environments. J Ambient Intell Smart Environ 2009;1(2):157–71.
[11] Hansen DW, Ji Q. In the eye of the beholder: a survey of models for eyes and gaze. IEEE Trans Pattern Anal Mach Intell 2010;32(3):478–500.
[12] Valenti R, Sebe N, Gevers T. Combining head pose and eye location information for gaze estimation. IEEE Trans Image Process 2011;21(2):802–15.
[13] Alberto Funes Mora K, Odobez JM. Geometric generative gaze estimation (G3E) for remote RGB-D cameras. In: Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition; 2014 Jun 23–28; Columbus, OH, USA. New York: IEEE; 2014. p. 1773–80.
[14] Zhang X, Sugano Y, Bulling A. Evaluation of appearance-based methods and implications for gaze-based applications. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems; 2019 May; Glasgow Scotland, UK. New York: Association for Computing Machinery; 2019. p. 1–13.
[15] Lu W, Li Y, Cheng Y, Meng D, Liang B, Zhou P. Early fault detection approach with deep architectures. IEEE Trans Instrum Meas 2018;67(7):1679–89.
[16] Zhang X, Sugano Y, Fritz M, Bulling A. MPIIGaze: real-world dataset and deep appearance-based gaze estimation. IEEE Trans Pattern Anal Mach Intell 2019;41(1):162–75.
[17] Yu Y, Liu G, Odobez JM. Deep multitask gaze estimation with a constrained landmark-gaze model. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018 Sep 9–14; Munich, Germany. New York: Springer; 2018. p. 456–74.
[18] Fischer T, Jin Chang H, Demiris Y. Rt-gene: real-time eye gaze estimation in natural environments. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018 Sep 9–14; Munich, Germany. New York: Springer; 2018. p. 334–52.
[19] Choe KW, Blake R, Lee SH. Pupil size dynamics during fixation impact the accuracy and precision of video-based gaze estimation. Vision Res 2016;118:48–59.
[20] Guestrin ED, Eizenman M. General theory of remote gaze estimation using the pupil center and corneal reflections. IEEE Trans Biomed Eng 2016;53(6):1124–33.
[21] Lu F, Sugano Y, Okabe T, Sato Y. Adaptive linear regression for appearancebased gaze estimation. IEEE Trans Pattern Anal Mach Intell 2014;36(10):2033–46.
[22] Huang MX, Kwok TC, Ngai G, Leong HV, Chan SC. Building a self-learning eye gaze model from user interaction data. In: Proceedings of the 22nd ACM international conference on Multimedia; 2014 Nov; Orlando Florida, USA. New York: Association for Computing Machinery; 2014. p. 1017–20.
[23] Liu G, Yu Y, Mora KAF, Odobez JM. A differential approach for gaze estimation. IEEE Trans Pattern Anal Mach Intell 2019;43(3):1092–9.
[24] Venturelli M, Borghi G, Vezzani R, Cucchiara R. From depth data to head pose estimation: a siamese approach. In: Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (Volume 5); 2017 Feb 27–Mar 1; Porto, Portugal. New York: Springer; 2017. p. 194–201.
[25] Liu G, Yu Y, Mora KAF, Odobez JM. A differential approach for gaze estimation with calibration. IEEE Trans Pattern Anal Mach Intell 2021;43(3):1092–9.
[26] Bromley J, Guyon I, Lecun Y, Säckinger E, Shah R. Signature verification using a "siamese" time delay neural network. In: Proceedings of the 6th International Conference on Neural Information Processing Systems; 1993 Nov. Denver, CO, USA: Morgan Kaufmann Publishers; 1994. p. 737–44.
[27] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. In: Proceedings of the 27th International Conference on Neural Information Processing Systems—Volume 1; 2014 Dec. Cambridge: MIT Press; 2014. p. 568–76.
[28] Simonyan K, Zisserman A. Very deep convolutional networks for largescale image recognition. In: Proceeding of International Conference on Learning Representations 2015; 2015 May 7–9; San Diego, CA, USA. New York: WikiCFP; 2015.
[29] Zhang X, Sugano Y, Fritz M, Bulling A. Appearance-based gaze estimation in the wild. In: Proceeding of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015 Jun 7–12; Boston, MA, USA. New York: IEEE; 2015. p. 4511–20.
[30] Lian D, Hu L, Luo W, Xu Y, Duan L, Yu J, et al. Multiview multitask gaze estimation with deep convolutional neural networks. IEEE Trans Neural Netw Learn Syst 2019;30(10):3010–23.
[31] Park S, Spurr A, Hilliges O. Deep pictorial gaze estimation. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018 Sep 9–14; Munich, Germany. New York: Springer; 2018. p. 741–57.
[32] Liu J, Francis BSL, Rajan D. Free-head appearance-based eye gaze estimation on mobile devices. In: Proceeding of 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC); 2019 Feb 11–13; Okinawa, Japan. New York: IEEE; 2019. p. 232–7.
[33] Wong ET, Yean S, Hu Q, Lee BS, Liu J, Deepu R. Gaze estimation using residual neural network. In: Proceeding of 2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops); 2019 Mar 11–15; Kyoto, Japan. New York: IEEE; 2019. p. 411–4.
[34] Dubey N, Ghosh S, Dhall A. Unsupervised learning of eye gaze representation from the web. In: Proceeding of 2019 International Joint Conference on Neural Networks (IJCNN); 2019 Jul 14–19; Budapest, Hungary. New York: IEEE; 2019. arXiv:1904.02459v1.
[35] Funes-Mora KA, Odobez JM. Gaze estimation in the 3D space using RGB-D sensors. Int J Comput Vis 2016;118(2):194–216.
[36] Funes Mora KA, Monay F, Odobez JM. Eyediap: a database for the development and evaluation of gaze estimation algorithms from RGB and RGB-D cameras. In: Proceeding of the Symposium on Eye Tracking Research and Applications; 2014 Mar 26–28; Florida, UF, USA. New York: Association for Computing Machinery; 2014. p. 255–8.
[37] Sugano Y, Fritz M, Andreas Bulling X, et al. It's written all over your face: fullface appearance-based gaze estimation. In: Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2017 Jul 21–26; Honolulu, HI, USA. New York: IEEE; 2017. p. 51–60.
[38] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Proceeding of the 25th International Conference on Neural Information Processing Systems—Volume 1; 2012 Dec 14–16; Siem Reap, Cambodia. LaneRed Hook: Curran Associates Inc; 2012. p. 1097–105.
[39] Ogusu R, Yamanaka T. Lpm: learnable pooling module for efficient full-face gaze estimation. In: Proceeding of 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019); 2019 May 14–18; Lille, France. New York: IEEE; 2019. p. 1–5.
[40] Sugano Y, Matsushita Y, Sato Y. Learning-by-synthesis for appearance-based 3D gaze estimation. In: Proceeding of 2014 IEEE Conference on Computer Vision and Pattern Recognition; 2014 Jun 23–28; Columbus, OH, USA. New York: IEEE; 2014. p. 1821–8.
[41] Zhang X, Sugano Y, Bulling A. Revisiting data normalization for appearancebased gaze estimation. In: Proceeding of the 2018 ACM Symposium on Eye Tracking Research & Applications. 2018 Jun 14–17; Warsaw, Poland. New York: Association for Computing Machinery; 2018. p. 1–9.
[42] Sugano Y, Matsushita Y, Sato Y, Koike H. An incremental learning method for unconstrained gaze estimation. In: Proceeding of European Conference on Computer Vision; 2008 Oct 12–18; Marseille, France. Berlin: Springer; 2008. p. 656–67.
[43] Zhang X, Huang MX, Sugano Y, Bulling A. Training person-specific gaze

estimators from user interactions with multiple devices. In: Proceeding of the 2018 CHI Conference onHumanFactors in Computing Systems;2018 Apr21–26;Montréal, QC, Canada. New York: Association for Computing Machinery; 2018. p. 624.
[44] Yu Y, Liu G, Odobez JM. Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In: Proceeding of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. New York: IEEE; 2019. p. 2019.
[45] Veges M, Varga V, L}orincz A, András L. 3D human pose estimation with Siamese equivariant embedding. Neurocomputing 2019;339:194–201.
[46] Doumanoglou A, Balntas V, Kouskouridas R, Kim TK. Siamese regression networks with efficient mid-level feature extraction for 3D object pose estimation. In: Proceeding of 29th Conference on Neural Information Processing Systems (NIPS 2016); 2016 Dec 5–10; Barcelona, Spain; 2016.
[47] Simo-Serra E, Trulls E, Ferraz L, Kokkinos I, Fua P, Moreno-Noguer F. Discriminative learning of deep convolutional feature point descriptors. In: Proceeding of 2015 IEEE International Conference on Computer Vision (ICCV); 2015 Dec 7–13; Santiago, Chile. New York: IEEE; 2015. p. 118–26.
[48] Wang J, Song Y, Leung T, Rosenberg C, Wang J, Philbin J, et al. Learning negrained image similarity with deep ranking. In: Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition; 2014 Jun 23–28; Columbus, OH, USA. Washington, DC: IEEE Computer Society; 2014. p. 1386–93.
[49] Baltrusaitis T, Robinson P, Morency LP. Continuous conditional neural fields for structured regression. In: Proceeding of European conference on computer vision; 2014 Sep 6–12; Zurich, Switzerland. Berlin: Springer; 2014. p. 593–608.
[50] Lepetit V, Moreno-Noguer F, Fua P. EPnP: an accurate o(n) solution to the PnP problem. Int J Comput Vis 2009;81(2):155–66.
[51] Schneider T, Schauerte B, Stiefelhagen R. Manifold alignment for person independent appearance-based gaze estimation. In: Proceeding of 2014 22nd International Conference on Pattern Recognition; 2014 Aug 24–28; Stockholm, Sweden. New York: IEEE; 2014. p. 1167–72.
[52] Mora KAF, Odobez JM. Person independent 3D gaze estimation from remote RGB-D cameras. In: Proceeding of 2013 IEEE International Conference on Image Processing. 2013 Sep 15–18; Melbourne, Australia. New York: IEEE; 2013. p. 2787–91.
[53] LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE 1998;86(11):2278–324.