Journal of Image and Signal Processing 图像与信号处理, 2023, 12(2), 144-157
Published Online April 2023 in Hans. https://www.hanspub.org/journal/jisp
https://doi.org/10.12677/jisp.2023.122015

Hans汉斯

基于预测一致性嵌入的注视目标检测

史俊彪，骆文杰，熊思璇，单东风，江朝晖，韩　超

合肥工业大学计算机科学与信息工程学院，安徽　合肥

收稿日期：2023年3月25日；录用日期：2023年4月15日；发布日期：2023年4月28日

摘　要

本文研究了第三人称视角下图像的注视目标检测问题我们提出了一个深度架构推断场景中的人在看哪里。该模型在蕴含丰富上下文信息的场景图像、深度图像和头部图像上进行训练。与现有的技术不同，我们的模型不需要监视注视角度，不依赖于头部方向信息和眼睛信息。大量的实验表明，我们的方法在多个基准数据集上具有更强的性能。我们还研究了注视目标检测的域自适应方法，使用一致性嵌入确保源域和目标域对齐，使得我们的模型能够有效地处理数据集之间的间隙。

关键词

注视目标检测，注视跟随，域自适应，RGB图像，深度图像

Gaze Target Detection Based on Predictive Consistency Embedding

Junbiao Shi, Wenjie Luo, Sixuan Xiong, Dongfeng Shan, Chaohui Jiang, Chao Han

School of Computer Science and Information Engineering, Hefei University of Technology, Hefei Anhui

Received: Mar. 25th, 2023; accepted: Apr. 15th, 2023; published: Apr. 28th, 2023

Abstract

In this paper, we study the problem of gaze target detection in images from the third person perspective. We propose a deep architecture to infer where people are looking in the scene. The model is trained on scene image, depth image and head image containing rich contextual information. Unlike existing technologies, our model does not need to monitor gaze angles and does not rely on head direction information and eye information. A large number of experiments show that our method has stronger performance on multiple benchmark data sets. We also study a domain adaptive approach to gaze target detection, using consistency embedding to ensure the alignment of source and target domains, so that our model can effectively deal with gaps between datasets.

文章引用: 史俊彪, 骆文杰, 熊思璇, 单东风, 江朝晖, 韩超. 基于预测一致性嵌入的注视目标检测[J]. 图像与信号处理, 2023, 12(2): 144-157. DOI: 10.12677/jisp.2023.122015

Keywords

Gaze Target Detection, Gaze Follow, Domain Adaptation, RGB Image, Depth Image

1. 引言

注视行为表明了一个人的视觉注意力，可以指明一个人对什么感兴趣，有助于破译和预测人们的互动、意图或行为[1] [2] [3]。人类有一种非凡的能力，可以察觉别人的目光方向，根据别人的目光来确定他们的注视目标，确定别人的注意力[4]。然而，自动执行和量化这些仍然是一个具有挑战性的问题。注视行为分析的研究分为注视估计和注视目标检测[1] [5] [6]。注视估计指的是确定人的注视方向(通常在3D中)，而不是准确定位场景中的人在看哪里[7]。相反，注视目标检测是推断场景中每个人(2D 或 3D)正在看哪里[1] [3] [8]。本文讨论了在第三人称视角下的二维图像中的注视目标检测，在这方面，以往工作提出了由两条路径组成的基于卷积神经网络架构。其中一条路径从全局图像中学习特征嵌入，另一条路径则对待测注视目标的人的头部图像进行建模[9] [10]。方法[9] [10]进行了空间建模，Chong 等人[5]通过对全局场景和头部图像随时间的嵌入进行建模(即应用时空建模)，扩展了上述的双路径架构。[5]方法相对于早期研究结果有所改善，但仍缺乏对场景深度的理解。因此沿着注视方向，不同深度有多个可能被注视的目标，就会导致错误预测。[1] [11]通过集成深度图像，解决这一缺陷。Fang 等人[1]还依赖于头部姿态检测，眼睛检测和眼睛特征提取。这样的模型虽然提高了注视目标检测精度，但是现实生活中可能容易出错，例如，当眼睛不可见或不可检测时。[11]使用辅助网络估计场景深度、此外还使用伪标签进行3d 注视方向估计。[11]的性能依赖于可靠的深度和方向伪标签。

我们不需要监督注视角度，这简化了我们的训练过程。与[5]不同的是，我们只应用了空间处理，但仍然能够检测到视频中的每一帧的注视目标。与[1] [11]类似，我们使用深度图像。使用三条路径来处理：1) 头部图像，2) 全局场景图像，3) 深度图像，深度图是通过独立的单目深度估计从 RGB 图像[12]中获得的。因为我们不需要检测头部姿态和眼睛位置，我们的计算成本更低，比[1]简单。与[11]不同的是我们的模型没有使用额外的模块来估计头部特征的注视方向。我们使用头部模块隐式的学习注视方向特征。本文研究了不同的方式对于注视目标估计任务的有效性。我们进行了全面的实验分析。不仅我们所提出的方法，而且它的一些变体也超过了现有方法的精度。

训练好的注视目标检测模型的泛化能力对其在实际中的应用至关重要。然而，实证分析表明，当在与训练数据集不同的数据集上的测试时，注视目标检测模型的性能显著降低。基于此，本文研究了域自适应问题，并提出了一种新的域自适应方法集成到所提出的注视目标检测模型中。显著提高了结果。我们研究的主要贡献可以概括为以下几点。

提出了一个新颖包含深度信息的注视目标检测模型，用于检测第三人称视角捕获的 2D 图像中的注视目标。我们在几个基准数据集上的实验证明了所提模型的性能优于当前的注视目标检测方法。

我们研究了注视目标检测的域自适应问题，并提出了预测一致性的目标域嵌入表示。设计了预测一致性损失，使其能够测量源域和目标域之间的位移。该方法提高了对目标域数据集的性能。

2. 相关工作

下面，我们描述了注视目标检测任务的相关研究。总结了域自适应的研究概况，重点介绍了域适应在注视行为分析和视觉数据的应用。

2.1. 注视目标检测

注视目标检测在多个领域都有应用，如人机交互系统、计算机视觉和机器人技术。其中重要的时理解感兴趣的对象，预测行动[13] [14]等等。现有的注视目标检测工作依赖于特定的传感器(眼动仪[3]，VR/AR 设备[15]，RGBD 摄像机[6] [16]等)或适用于特定设置(面对面会议)或用于注视行为分析(相互注视[17]，共同注视[18] [19])。另一种分类时关于目标在二维图像[4] [5] [9] [10] [20]还是三维空间[6] [16] [21] [22]。

本文主要研究在无约束环境下第三人称视角下采集的二维单幅图像的注视目标检测问题。[10]是最早的注视目标检测深度学习架构之一，它呈现双路径架构。一个分支获取场景图像来估计显著性(所谓的显著性路径)，另一个分支(所谓的注视路径)使用头部图像作为输入对注视方向建模。并且将头部位置信息注入注视路径，显著提高注视目标检测结果。后面的工作[4]采用了上述的双通路架构，其他工作[1] [5] [9] [11] [20]采用了双通路架构集合头部信息注入。不同的是 Chong 等人[4]扩展了[10]的架构，通过同时学习注视角度和显著性来检测不在场景中的注视目标(所谓的帧外注视目标)。Chong [4]还集成了 CNN-LSTM，处理注视和显著性路径的特征嵌入，学习时间上的注视行为。尽管[5] [9] [10] [20]提出的模型能够有效的处理注视目标检测任务，但是他们都未能解决场景深度对注视目标检测任务的影响。例如，在注视方向上有多个物体，它们的场景深度不同，这些方法很难确定正确的注视目标。针对于场景深度问题，[1]利用深度信息和三维注视去除场景深度的影响。然而[1]的计算量很大，因为它需要检测头部姿势和眼睛。Jin 等人[11]的模型也包括深度场景信息，他们设计了和[5]一样的预测注视的主要网络。此外为了提高性能，加入场景深度信息，他们引入了两个辅助网络，一个用于学习场景深度特征，另一个用于学习三维注视方向特征。

我们的工作在几个方向与现有技术不同。首先，与[9] [10] [20]不同，我们不需要监督凝视角度。与[5] [9] [10] [20]不同的是，我们采用单目深度估计方法[12]获得深度图像，通过处理人的相对深度改善空间建模。与[5]不同的是，我们不仅利用头部特征调节场景信息，还使用头部特征调节深度信息，从而提高性能。另外，我们只依赖于空间信息，没有进行时空数据的处理。我们提出了一种三路径网络，同时学习头部特征、场景特征和深度特征，并使用头部特征对场景特征和深度特征调节。不使用[1]中应用的头部姿势和眼睛图像。与[11]不同的是，我们的工作不需要额外的深度和方向伪标签，也不需要额外的网络来显示的学习 3D 注视方向。

与现有技术相比，我们的框架实现了更好的性能，甚至超过了人类的性能。重要的是，针对于二维图像注视目标检测任务，我们提出了简单但有效的域自适应方法，以提高网络的泛化性。

2.2. 域自适应

无监督域自适应(UDA)，是一种被广泛研究的方法，用于处理训练数据和测试数据属于不同分布时由于域间隙产生的问题。UDA 将仅使用源域的监督训练模型推广到缺乏标签的目标域。关于 UDA 可以分为三部分 1) 基于差异的技术[23] [24] [25]，试图在特征级别上最小化源域和目标域之间的距离。2) 对抗方法[26] [27]，具有生成器和鉴别器，并试图让生成器创建的特征尽可能接近源域。并且使用鉴别器来促进域混淆。3) 利用自监督学习[28] [29] [30]减少域位移。

针对注视估计任务的域自适应相对较少，对于注视目标检测问题，利用添加梯度反转的域分类器，

混淆源域和目标域。以及应用 RGB→Depth 和 Depth→RGB 模态转换。对于注视估计问题 Kell [31]采用[32]的对抗性判别域自适应，其中判别器识别图像特征为二值分类任务，并根据注视估计任务的左右对称性，通过计算原始图像和水平翻转图像的注视，最小化两者之间的角度差。Yu [33]从注视重定向的监督解决注视域自适应问题，因为不同的人的眼睛结构会导致域间隙，从而导致表现不佳。为了解决问题 Yu [33]从现有的参考样本中生成合成眼图像，并定义注视重定向损失以及循环一致性损失。

然而这些方法大部分都是针对于注视估计的域自适应问题，针对于注视目标检测域自适应问题，我们在注视目标检测模块和我们定义的域分类器之间注入一个梯度反转层。我们还引入预测一致性的嵌入来消除源和目标域之间的间隙。

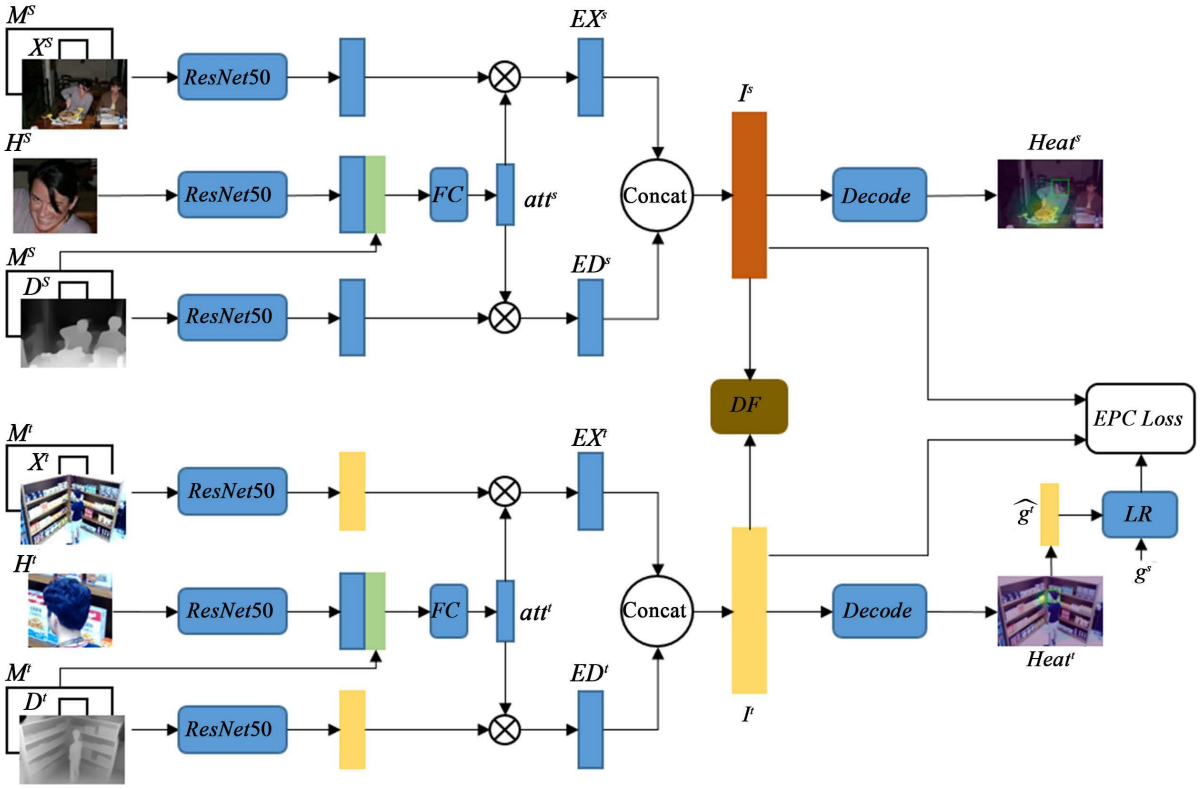

Figure 1. Domain adaptive gaze target detection model architecture
图 1. 域自适应注视目标检测模型架构

3. 方法

模型的整体架构如图 1 所示。我们用上标 s 表示为源域，上标 t 表示为目标域。我们的模型将源域场景 RGB 图像 X_i^s、深度图像 D_i^s、头部图像 H_i^s、头部位置掩码 M_i^s 和目标域场景 RGB 图像 X_j^s、深度图像 D_j^s、头部图像 X_j^s、头部位置掩码 M_j^s 作为输入。其中深度图像 D 是通过对场景 RGB 图像 X 使用最先进的单目深度估计器[12]得到。头部位置掩码 M 通过对头部位置的像素值设置为 1，其余位置像素值设置为 0 得到。模型输出两个和场景图像相同大小的注视热图，源域注视热图 $Heat_i^s$，目标域注视热图 $Heat_j^t$，注视热图中像素值最大的像素点位置为预测的注视位置。除了预测注视热图，我们的模型还输出 $InOut$，$InOut$ 表示注视目标在帧内的概率，当注视目标在帧内时 $InOut = 1$。

我们模型在 Chong [5]的基础上，加入了深度信息。它与 RGB 图像一起提供了更加丰富的场景信息。

与 Chong [5]不同的是，我们还加入了域自适应模块，包括添加梯度反转[34]的域分类器和预测一致性嵌入。下面我们将模型分为两部分介绍：注视目标检测模块和域自适应模块。

3.1. 注视目标检测模块

该模块由三个子模块组成：头部模块(*HN*)用于处理头部图像，并生成头部调节特征。场景模块(*XN*)用于处理场景图像 X。深度模块(*DN*)用于处理深度图像 D。预测模块(*PN*)，用于生成注视热图和注视目标是否在帧内的判断。

头部模块：给定 RGB 场景图像 X_i^s，我们对目标人物头部进行裁剪得到头部图像 H_i^s。使用 ResNet-50 后面跟一个平均池化层，对头部图像进行特征提取，得到头部特征 EH_i^s，同时使用三个最大池化层处理头部位置掩码 M_i^s，拉伸之后与头部特征 EH_i^s 拼接。拼接特征经过一个全连接层得到头部调节特征 att_i^s。

$$att_i^s = FC\left(Concat\left(ReaNet\left(H_i^s\right), MaxPool\left(M_i^s\right)\right)\right) \tag{1}$$

场景模块：场景模块的主干结构与头部模块相同为 ResNet-50。该模块以场景图像 X_i^s 和头部位置掩码 M_i^s 的拼接作为输入提取场景特征。并且将场景特征的每一个通道乘以头部模块生成的头部调节特征 att_i^s，得到加权的场景特征 EX_i^s。

$$EX_i^s = ResNet\left(Concat\left(X_i^s, M_i^s\right)\right) \otimes att_i^s \tag{2}$$

其中 $\otimes$ 为通道乘法。我们通过将场景特征和头部调节特征相乘，使得网络关注头部朝向场景中的物体上，这与[1] [5]一致。

深度模块：深度模块具有和场景模块相同的主干架构和相同的维度输入。该模块以深度图 D_i^s 和头部位置掩码 M_i^s 的拼接作为输入提取深度特征。和场景模块一样，我们将深度特征的每一个通道乘以头部模块生成的头部调节特征 att_i^s，得到加权的深度特征 ED_i^s。

$$ED_i^s = ResNet\left(Concat\left(D_i^s, M_i^s\right)\right) \otimes att_i^s \tag{3}$$

其中 $\otimes$ 为通道乘法。我们通过将深度特征和头部调节特征相乘，通过头部特征包含的头部姿态信息，去除场景深度影响。

预测模块：我们将场景特征 EX_i^s 和深度特征 ED_i^s 的通道和输入到编码器 *Encode*，得到最终的注视特征 I_i^s。预测模块输出 2D 注视热图 $Heat_i^s$ 和注视目标在帧内的概率 $InOut_i^s$。为了获得 2D 注视热图，我们使用多层解码器 *Decode*，它将最终的注视特征作为输入：

$$Heat_i^s = Decode\left(Encode\left(EX_i^s \oplus ED_i^s\right)\right) \tag{4}$$

此外我们还将最终的注视特征 EG_i^s 经过三个全连接层计算注视目标在帧内的概率 $InOut_i^s$：

$$InOut_i^s = FC\left(Encode\left(EX_i^s \oplus ED_i^s\right)\right) \tag{5}$$

3.2. 域自适应模块

域自适应模块由两个部分组成。1) 域分类器：它决定注视特征是属于源域还是目标域。2) 一致性嵌入：它使得源域和目标域的注视特征对齐，以提高注视目标检测的性能。

注视目标检测模块权重对于源域和目标域共享。经过注视目标检测模块，源域和目标域得到对应的注视热图 $Heat_i^s$ 和 $Heat_j^t$。注视热图最大像素值位置为预测的注视目标位置。我们使用 $\widehat{g_i^s}$ 表示源域第 i 个样本的预测注视目标位置，其对应的注视特征为 I_i^s，对应的注视位置标签为 g_i^s。使用 $\widehat{g_j^t}$ 表示目标域

第 j 个样本的预测注视目标位置，其对应的注视特征为 I_j^t。我们通过源域和目标域的预测注视目标位置约束注视特征，使得源域和目标域的注视特征对齐。

域分类器：域分类器在源域和目标域之间执行二分类，我们使用注视特征作为域分类器的输入。域分类器通过梯度反转层[34]连接在注视目标检测模块上。梯度反转层在反向传播训练时将梯度乘以某个常负数。梯度反转层确保学习到的特征尽可能不区分源域和目标域。

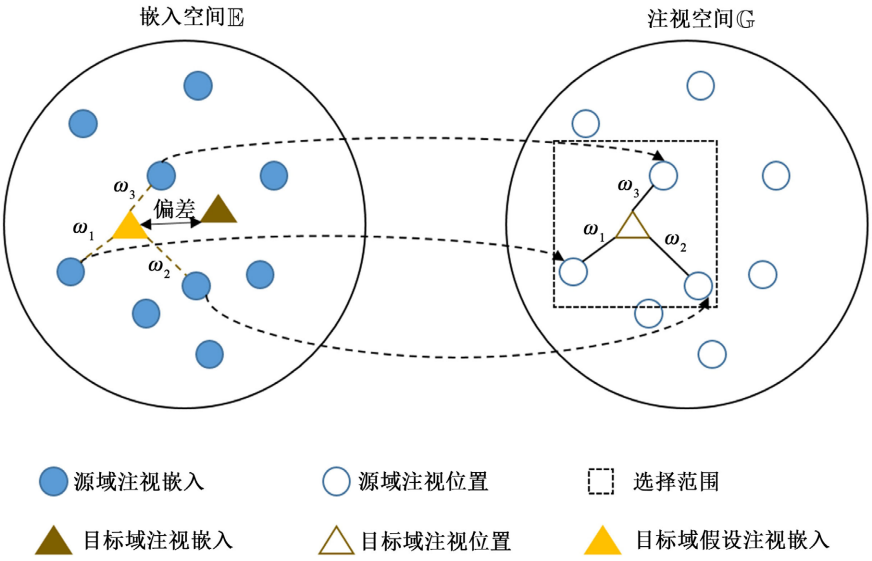

Figure 2. Embedding with prediction consistency
图 2. 预测一致性嵌入

一致性嵌入：将注视特征所在的空间描述为嵌入空间 $\mathbb{E}$，注视位置所在的标签描述为注视空间 $\mathbb{G}$。我们提出了一种域注视的局部线性表示方法(LR)，对于目标域，利用源于注视位置线性表示目标域注视位置。对于目标域中的每一个样本，注视目标检测模块预测得到的注视位置为假设标签。我们在注视空间组合 K 个源域注视位置标签来描述它。

我们首先在注视空间中定义每一个目标域预测注视位置的领域，只有当源域注视位置标签 g_i^s 与目标域预测注视位置 $\widehat{g_j^t}$ 差异小于 μ 时，g_i^s 被视为目标域预测注视位置 $\widehat{g_j^t}$ 的邻居。我们用 N_j 来描述 $\widehat{g_j^t}$ 的所有邻居集合，定义为：

$$N_j = \left\{ g_i^s \mid \max\left(\left| x_i^s - \widehat{x_j^t} \right|, \left| y_i^s - \widehat{y_j^t} \right| \right) < \mu \right\} \tag{6}$$

在拥有 K 个以上的领域，每一个目标域预测注视位置 $\widehat{g_j^t}$ 随机抽取 K 个生成 N_j，用于表示 $\widehat{g_j^t}$。我们定义权重 ω_{ij} 表示第 i 个源域注视位置对第 j 个目标域预测注视位置 $\widehat{g_j^t}$ 重构的贡献，我们的目的时为每一个 ω_{ij} 找到最合适的解。

对于与 2D 注视位置，邻域数量越大，意味着在训练过程中很难找到一个合适的解来最小化重构损失 $E(\omega)$。我们引入 L2 正则化确保唯一解。重构损失 $E(\omega)$ 表示为：

$$E(W_j) = \left\| \widehat{g_j^t} - \sum_{i=1}^{k} \omega_{j,i} \cdot g_i^s \right\|_2^2 + \lambda \sum_{i=1}^{k} \omega_{j,i}^2, g_i^s \in N_j, \sum_{i=1}^{k} \omega_{j,i} = 1 \tag{7}$$

其中 $W_j = \left[\omega_{j,1}, \omega_{j,2}, \cdots, \omega_{j,k} \right]$，可以写成矩阵形式：

$$E\left(W_j\right)=W_j^{\mathrm{T}}\left(\widehat{G_j^t}-G_i^s\right)^{\mathrm{T}}\left(\widehat{G_j^t}-G_i^s\right)W_j+\lambda W_j^{\mathrm{T}}W_j=W_j^{\mathrm{T}}\left(P_j+\lambda I\right)W_j \tag{8}$$

其中 $\widehat{G_j^t}=\left[\widehat{g_j^t},\cdots,\widehat{g_j^t}\right]_{1\times k}$，$G_i^s=\left[g_i^s,\cdots,g_k^s\right]$，$P_j$ 为局部协方差矩阵，定义为：

$$P_j=\left(\widehat{G_j^t}-G_i^s\right)^{\mathrm{T}}\left(\widehat{G_j^t}-G_i^s\right) \tag{9}$$

由拉格朗日乘子法得到使重构损失 $E\left(W_j\right)$ 最小的解 W_j^* 为：

$$W_j^*=\frac{\left(P_j+\lambda I\right)^{-1}1_k}{1_k^{\mathrm{T}}\left(P_j+\lambda I\right)^{-1}1_k} \tag{10}$$

当最优权重 $W_j^*=\left[\omega_{j,1}^*,\cdots,\omega_{j,k}^*\right]$ 时，LR 的形式表示为：

$$\widehat{g_j^t}=\sum_{i=1}^k\omega_{j,i}^*g_i^s,\ g_i^s\in N_j \tag{11}$$

我们将注视空间 $\mathbb{G}$ 中的线性关系转移到嵌入空间 $\mathbb{E}$ 中，生成目标域假设注视嵌入。对于目标域注视嵌入 I_j^t，对应的假设嵌入表示为：

$$\widehat{I_j^t}=\sum_{i=1}^k\omega_{j,i}^*I_i^s \tag{12}$$

如图 2 所示，注视空间 $\mathbb{G}$ 中的 LR 权值被继承到嵌入空间 $\mathbb{E}$。对于嵌入空间$\mathbb{E}$中的目标域注视嵌入，我们通过公式生成其对应的假设嵌入 $\widehat{I^t}$。目标域的假设嵌入和源域嵌入之间的线性关系与目标域预测注视位置和源域注视位置之间的线性关系相同。

3.3. 损失函数

我们首先使用源域训练我们的注视目标检测模块，注视目标检测模块的总损失是注视热图上的均方差损失 $L_{heatmap}$ 和注视目标是否在帧内的二进制交叉熵损失 $L_{in/out}$ 的加权和。

$$L_{gaze}=\lambda_{haetmap}L_{heatmap}+\lambda_{in/out}L_{in/out} \tag{13}$$

其中 $\lambda_{haetmap}$ 和 $\lambda_{in/out}$ 为其权重参数。

对于注视热图损失，我们在标签中给出的注视位置周围放置高斯权重，作为注视热图标签。并使用均方差计算注视热图损失。给定训练过程中的 Bt 大小批次的源样本，损失函数如下：

$$L_{heatmap}=\frac{1}{Bt}\sum_{i=1}^{Bt}MSE\left(Heat_i,Heat_i^{gt}\right) \tag{14}$$

其中 $Heat_i$ 是该批次中第 i 个样本，$Heat_i^{gt}$ 为其对应标签，MSE 表示均方差损失函数。

我们通过交叉熵损失函数计算注视目标落在帧内还是帧外的损失，损失函数如下：

$$L_{in/out}=-\frac{1}{Bt}\sum_{i=1}^{Bt}InOut_i^{gt}\cdot\log\left(InOut_i\right)+\left(i-InOut_i^{gt}\right)\cdot\log\left(1-InOut_i\right) \tag{15}$$

为了实现源域和目标域之间的域自适应，我们提出了两个 DA 损失，其中预测一致性损失 L_{EPC} 计算源域和目标域之间的偏差，它确保相源域和目标域相同注视位置的样本具有相同的嵌入特征。同时我们以对抗损失 L_D 缓解域偏移，使得目标域嵌入更接近源域嵌入。

对抗损失函数 L_D 如下：

$$L_D=\frac{1}{Bt}\left(\sum_{i=1}^{Bt}\log\left(DF\left(I_i^s\right)\right)+\sum_{i=1}^{Bt}\log\left(1-DF\left(I_i^t\right)\right)\right) \tag{16}$$

预测一致性损失函数 L_{EPC} 如下：

$$L_{EPC}=\frac{1}{Bt}\sum_{j=1}^{bt}d\left(I_j^t,\sum_{i=1}^{k}\omega_{ji}^{*}\cdot I_i^s\right) \tag{17}$$

其中函数 d 为 L1 距离。L_{EPC} 测量目标域假设嵌入与预测嵌入之间的距离。此外由于目标域假设嵌入是源域嵌入的线性组合，L_{EPC} 还评估了源域和目标域之间的偏差。在训练过程中，随着目标域假设嵌入与预测嵌入越来越近，域之间的偏移逐渐消除。

因为有些数据集没有 *InOut* 标签，所以我们在进行带有域自适应的注视目标检测任务的时候去除了 $L_{in/out}$ 损失。应用我们域自适应模块的总损失为 $L_{heatmap}$、L_D 和 L_{EPC} 三个损失。域自适应注视目标检测总损失函数如下：

$$L_{total}=\lambda_{heatmap}L_{heatmap}+\lambda_D L_D+\lambda_{EPC}L_{EPC} \tag{18}$$

其中 $\lambda_{heatmap}$、λ_D 和 λ_{EPC} 为其权重参数。

4. 实验

4.1. 数据集、实现细节、评价指标

数据集：GazeFollow [10]数据集包含 122,143 张图片，包含 130,339 个头部位置和相应的注视点标签，测试集包含 4782 个标签，其余用于训练集。因为 GazeFollow 数据集是静态图片，并且不包含注视点落在帧内还是帧外的标签。VideoAttentionTarget [5]数据集由来自 Youtube 上各种来源的 1331 个视频剪辑组成。VideoAttentionTarget 的注释包括 164,541 个帧级头部边界框，109,574 个帧内凝视目标，54,967 个帧外凝视。测试集包含 31,978 个标签，其余用于训练集。GOO [35]数据集是包含 24 类杂货的货架图像集合。在每张图片中，一个人在看着货架上的一件物品，GOO 数据集是注视目标检测任务中第一个同时使用真实图片和合成图片的数据集。在本文中，我们只使用了真实图片，没有使用合成图片。训练集包含 9552 张图片，测试集包含 2156 张图片。

实现细节：我们在 PyTorch 中实现我们的模型。场景图片、头部图片和深度图片都被标准化并调整为 224 × 224 场景模块、头部模块和深度模块主干基于 ResNet-50。场景主干在 Place 数据集[36]上进行预训练，头部主干在 Eyediap 数据集[37]上进行预训练，深度主干也在 Place 数据集[36]上进行预训练。

由于注视目标检测模块决定 λ_{EPC} 中的邻域和线性注视表示，因此需要一个训练好的模型来生成可靠的目标域预测注视位置。我们首先使用源域数据集对注视目标检测模块进行 70 期的预训练，批次大小为 16，学习率为 2.5×10^{-4}。

对于带域自适应的注视目标检测任务，我们要同时优化 $L_{heatmap}$、L_D 和 L_{EPC}，进行了 70 次迭代，批次大小为 16，学习率为 0.0001。我们首先更新源域和目标域的注视嵌入 I^s，I^t。并通过域分类器 *DF* 计算对抗损失 L_D。然后更新目标域预测注视位置 $\widehat{g^t}$，同时预测源域的预测注视位置。对于目标域预测注视位置 $\widehat{g^t}$，我们构造 N_j 并计算 L_{EPC}。我们使用源域标签进行 LR 操作，以获得更高的精度。通过反向传播对网络参数进行跟新，使得 $L_{heatmap}$、L_D 和 L_{EPC} 最小。

对于不带域自适应的注视目标检测任务，我们要同时优化 $\lambda_{heatmap}$ 和 $\lambda_{in/out}$。该模块在 GazeFollow 数据集上从头开始训练 70 期，批次大小为 16，学习率为 2.5×10^{-4}。拟合之后，我们在 VideoAttentionTarget 数据集上对模型进行微调。此外，我们在 GOO 数据集上从头开始训练我们的注视目标检测模块，训练 70 期，批次大小为 16，学习率为 2.5×10^{-4}。

评价指标：我们采用以下的指标来评估模型的性能。AUC↑：我们使用[38]提出的曲线下面积(AUC)标准来评估预测热图的置信度。↑表示 AUC 值越大精度越高。Dist↓：我们评估热图中最大值像素给出的

预测值与真实注视点标签之间的距离，↓表示距离越小注视点估计结果更准确。

4.2. 实验分析

4.2.1. 注视目标检测

Table 1. Evaluation on benchmark datasets.
表 1. 对基准数据集的评估

	GazeFollow [10]		VideoAttentionTarget [5]		GOO [35]	
	AUC	Dist	AUC	Dist	AUC	Dist
Random	50.4	0.484	50.5	0.458	-	-
Center	63.3	0.313	-	-	-	-
Fixed Bias	67.4	0.306	72.8	0.326	-	-
Recasens (2015) [10]	87.8	0.190	-	-	85.0	0.220
Chong (2018) [4]	89.6	0.187	83.0	0.193	-	-
Lian (2018) [9]	90.6	0.145	-	-	84.0	0.321
Chong (2020) [5]	92.1	0.137	86.0	0.134	79.6	0.252
Fang (2021) [1]	92.2	0.124	90.5	0.108	-	-
Jin (2022) [11]	92.0	0.118	90.1	0.116	-	-
Ours	92.8	0.132	94.3	0.124	92.0	0.153
人类表现	92.4	0.096	92.1	0.051	-	-

我们将我们的注视目标检测模块与表 1 中的以往方法进行比较。这些比较包括标准注视分析基线，1) 随机，2) 中心偏差，3) 固定偏差，其结果取自[10]。随机，表示通过从高斯分布中采样值来生成每一个像素的热图。中心偏差，预测结果总在图像的中心，固定偏差，预测是由训练集中与测试集中头部图像相似位置的头部位置的平均值。

我们的方法获得了同类方法中最好的结果。它甚至超过了 GazeFollow 和 VideoAttentionTarget 数据集中的人类表现，分别增加了 0.4%和 2.2%。在 VideoAttentiontarget 和 GOO 数据集上分别性能提升最显著，AUC 分别提升了 3.8%~11.3%和 2.0%~12.4%。在 Dist 方面我们的方法落后 Fang [1]和 Jin [11]，但性能优于其他方法。因为 Fang 方法与我们的模型相比，通过添加 1) 提取头部姿势，2) 检测眼睛，3) 提取眼睛特征的组件，呈现出更复杂的模型，这在显示应用中可能无法正确执行。另一方面，对于 Jin [11]，我们认为他们模型中的提取 3D 注视方向和生成深度图像的辅助网络有助于提高 Dist，而该方法在 AUC 方面比我们和其他的方法表现差。

此外，可以观察到，即使是空间模型，我们的方法优于 Chong [5]，其中包括 CONV-LSTM 网络。因此我们得出结论，与依赖 RGB 的视频相比，集成由 RGB 图像生成的深度图像可以获得更好的注视目标检测性能。

4.2.2. 跨数据集注视目标检测

本节研究了图像中注视目标检测任务的域自适应问题。为此我们在一个数据集上训练 Chong [5]的模型和提出的方法，而在一个完全不同的数据集上去测试训练的模型，相应的结果如表 2 所示。可以看出，我们的方法在所有跨域分析中都优于 Chong [5]。因此我们认为，我们添加的域自适应模块是有效的。

Table 2. Evaluation of gaze target detection performance across datasets.
表 2. 跨数据集注视目标检测评估

	训练集	测试集	AUC	Dist
Chong (2015) [5]	GazeFollow	GOO	77.3	0.270
Ours	GazeFollow	GOO	82.5	0.201
Chong (2015) [5]	VideoAttentionTarget	GOO	68.2	0.311
Ours	VideoAttentionTarget	GOO	86.1	0.184
Chong (2015) [5]	GOO	GazeFollow	62.5	0.410
Ours	GOO	GazeFollow	78.4	0.287
Chong (2015) [5]	GOO	VideoAttentionTarget	55.1	0.458
Ours	GOO	VideoAttentionTarget	70.3	0.312

4.2.3. 消融实验

为了更好的研究我们模型不同组成部分的贡献，我们训练了注视目标检测模块的以下变体。1) 仅包含场景模块：我们去除了头部模块和深度模块。头部特征没有和场景特征关联，仅仅通过头部位置掩码生成头部调节特征。2) 场景模块和头部模块：我们去除深度模块，场景模块和头部模块得到保留。场景模块的输入是场景图像和头部位置掩码的拼接。3) 头部模块和深度模块：我们去除场景模块，头部模块和深度模块得到保留，深度模块的输入是深度图像和头部位置掩码的拼接。4) 我们将深度模块移除，将场景模块的输入换为场景图像和深度图像的拼接。5) 我们将深度模块移除，将场景模块的输入换为场景图像、深度图像和头部位置掩码的拼接。6) 我们将公式 4 中提出的求和操作替换为拼接操作，拼接特征应用于解码器。相应的结果如表 3 所示。

结果表明，我们的注视目标检测模块的所有组成部分对于实现最佳性能都很重要。实验 1、2、3 和 4 让我们分别了解头部、场景和深度模块的贡献。其中最重要的是头部模块(将 AUC 提高了 16.9)，它提供了场景中人的头部方向信息，使得场景模块和深度模块更多的关注头部朝向的区域。贡献第二大的是场景模块(与 GazeFollow 数据集上的实验 3 相比，将 AUC 提高了 3.5, Dist 降低了 0.035).。使用没有深度模块的模型仍然无法处理不同深度的目标。我们提出的方法相比于实验 2，AUC 提高了 0.4，Dist 减少了 0.09。在 VideoAttentionTarget 数据集上的性能提升更高，AUC 提高了 2.5，Dist 减少了 0.01。

Table 3. Gaze target detection module ablation experiment
表 3. 注视目标检测模块消融实验

	Gaze Follow [10]		VideoAttentionTarget [5]	
	AUC	Dist	AUC	Dist
1	75.9	0.258	81.3	0.193
2	92.2	0.143	91.8	0.144
3	89.4	0.167	90.8	0.151
4	91.2	0.158	92.6	0.137
5	91.5	0.152	93.5	0.135
6	92.8	0.142	91.9	0.137
Ours	92.8	0.132	94.3	0.124

我们对于添加的域自适应模块也进行了消融实验。我们分别在不包含域自适应模块、仅包含 L_D 损失、仅包含 L_{EPC} 损失和我们完整的模型上进行实验。

Table 4. Ablation experiment of domain adaptation methods for gaze target detection
表 4. 注视目标检测域自适应方法消融实验

	训练集	测试集	AUC	Dist
Ours (w/o DA)	GazeFollow	GOO	78.5	0.281
Ours (L_D)	GazeFollow	GOO	80.3	0.269
Ours (L_{EPC})	GazeFollow	GOO	81.4	0.237
Ours (full)	GazeFollow	GOO	82.5	0.201
Ours (w/o DA)	VideoAttentionTarget	GOO	69.5	0.272
Ours (L_D)	VideoAttentionTarget	GOO	80.6	0.217
Ours (L_{EPC})	VideoAttentionTarget	GOO	84.7	0.198
Ours (full)	VideoAttentionTarget	GOO	86.1	0.184
Ours (w/o DA)	GOO	GazeFollow	63.4	0.386
Ours (L_D)	GOO	GazeFollow	70.8	0.339
Ours (L_{EPC})	GOO	GazeFollow	73.7	0.305
Ours (full)	GOO	GazeFollow	78.4	0.287
Ours (w/o DA)	GOO	VideoAttentionTarget	57.9	0.421
Ours (L_D)	GOO	VideoAttentionTarget	62.8	0.367
Ours (L_{EPC})	GOO	VideoAttentionTarget	68.1	0.349
Ours (full)	GOO	VideoAttentionTarget	70.3	0.312

表 4 展示了消融实验结果。实验结果表明我们的域自适应模块每一个损失函数都是重要的，同时使用它们可以获得最佳的性能。其中预测一致性嵌入对性能提升最大。

5. 可视化结果

图 3 展示了在 GazeFollow 数据集上的结果，其中我们展示了我们的注视目标检测模块(即没有域自适应)和 Chong [5]的结果以及标签。我们可以看到我们的模型可以有效地处理场景深度对注视目标检测任务的影响。在图 4 中我们展示了模型有和没有域自适应的可视化结果。源数据集为 GazeFollow，目标数据集为 GOO 时。从图 3 中可以观察到我们的域自适应方法和没有域自适应方法相比，显著提高了注视目标检测结果。

Figure 3. Visual results of gaze target detection module
图 3. 注视目标检测模块可视化结果

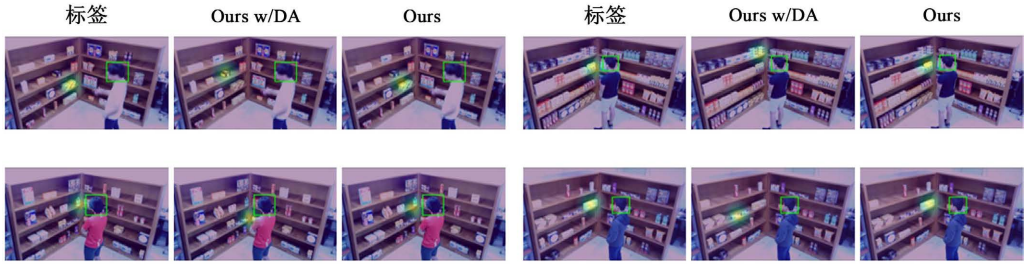

Figure 4. Visual results of domain adaptive module
图 4. 域自适应模块可视化结果

6. 结束语

我们提出了一种新颖的包含深度信息的注视目标检测模型，在第三人称的视角检测图像中的人在看哪里。我们的模型由三路径组成，1) 头部图像，2) 场景图像，3) 深度图像。场景图像和深度图像提供场景信息和深度信息。它与现有技术不同，因为它不依赖于注视角度的监督，不需要明确的头部方向信息或被检测目标的眼睛位置。大量的评估表明，所提出的方法优于现有方法。

本文还研究了用于注视目标检测的域自适应方法。为此我们想前面所描述的注视目标检测模块添加了域自适应模块。域分类器确保了学习到的特征尽可能不区分源域和目标域，一致性嵌入使得源域和目标域的注视特征对齐。我们的方法增强了目标数据集上的性能。在本文中我们没有从注视目标在帧内还是帧外的精度方面评估我们的方法。这是由于在多个数据集中缺乏相应的标签。

基金项目

安徽省重点研究与开发计划(202004d07020004)，安徽省自然科学基金项目(2108085MF203)，中央高校基本科研业务费专项资金(PA2021GDSK0072, JZ2021HGQA0219)。

参考文献

[1] Fang, Y., Tang, J.P., Shen, W., Shen, W., Gu, X., Song, L. and Zhai, G.T. (2021) Dual Attention Guided Gaze Target Detection in the Wild. *Proceedings of* 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, 20-25 June 2021, 11390-11399. https://doi.org/10.1109/CVPR46437.2021.01123

[2] Niewiadomski, R., Chauvigne, L., Mancini, M. and Camurri, A. (2018) Towards a Model of Nonverbal Leadership in Unstructured Joint Physical Activity. *Proceedings of the* 5*th International Conference on Movement and Computing* (*MOCO*'18), Genoa, 28-30 June 2018, 1-8. https://doi.org/10.1145/3212721.3212816

[3] Thakur, S.K., Beyan, C., Morerio, P. and Del Bue, A. (2021) Predicting Gaze from Egocentric Social Interaction Videos and IMU Data. *Proceedings of* 2021 *International Conference on Multimodal Interaction* (*ICMI*' 21), Montreal, 18-22 October 2021, 717-722. https://doi.org/10.1145/3462244.3479954

[4] Chong, E.J., Ruiz, N., Wang, Y.X., Zhang, Y., Rozga, A. and Rehg, J.M. (2018) Connecting Gaze, Scene and Attention: Generalized Attention Estimation via Joint Modeling of Gaze and Scene Saliency. *ECCV* 2018: 15*th European Conference*, Munich, 8-14 September 2018, 397-412. https://doi.org/10.1007/978-3-030-01228-1_24

[5] Chong, E.J., Wang, Y.X., Ruiz, N. and Rehg, J.M. (2020) Detecting Attended Visual Targets in Video. *Proceedings of* 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 5396-5406. https://doi.org/10.1109/CVPR42600.2020.00544

[6] Hu, Z., Yang, D., Cheng, S., Zhou, L., Wu, S. and Liu, J. (2022) We Know Where They Are Looking at From the RGB-D Camera: Gaze Following in 3D. *IEEE Transactions on Instrumentation and Measurement*, **17**, 1-14. https://doi.org/10.1109/TIM.2022.3160534

[7] Zhang, X., Huang, M.X., Sugano, Y. and Bulling, A. (2018) Training Person-Specific Gaze Estimators from User Interactions with Multiple Devices. *Proceedings of* 2018 *CHI Conference on Human Factors in Computing Systems*, Montreal, 21-26 April 2018, 1-12. https://doi.org/10.1145/3173574.3174198

[8] Liu, M., Li, Y. and Liu, H. (2020) 3D Gaze Estimation for Head-Mounted Eye Tracking System with Auto-Calibration Method. *IEEE Access*, **8**, 104207-104215. https://doi.org/10.1109/ACCESS.2020.2999633

[9] Lian, D., Yu, Z. and Gao, S. (2018) Believe It or Not, We Know What You Are Looking At! *ACCV* 2018: 14*th Asian Conference on Computer Vision*, Perth, 2-6 December 2018, 35-50. https://doi.org/10.1007/978-3-030-20893-6_3
Lian D, Yu Z, Gao S. Believe it or not, we know what you are looking at![C]//Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14. Springer International Publishing, 2019: 35-50.

[10] Recasens, A., Khosla, A., Vondrick, C. and Torralba, A. (2015) Where Are They Looking? *Advances in Neural Information Processing Systems*, **28**, 199-207.

[11] Jin, T., Yu, Q., Zhu, S., Lin, Z., Ren, J., Zhou, Y. and Song, W. (2022) Depth-Aware Gaze-Following via Auxiliary Networks for Robotics. *Engineering Applications of Artificial Intelligence*, **113**, Article ID: 104924. https://doi.org/10.1016/j.engappai.2022.104924

[12] Ranftl, R., Lasinger, K., Hafner, D., Schindler, K. and Koltun, V. (2020) Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**, 1-14.

[13] Li, Y., Liu, M. and Rehg, J. (2021) In the Eye of the Beholder: Gaze and Actions in First Person Video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2021.3051319

[14] Min, K. and Corso, J.J. (2021) Integrating Human Gaze into Attention for Egocentric Activity Recognition. *Proceedings of* 2021 *IEEE Winter Conference on Applications of Computer Vision* (*WACV*), Waikoloa, 3-8 January 2021, 1069-1078. https://doi.org/10.1109/WACV48630.2021.00111

[15] Dohan, M. and Mu, M. (2019) Understanding User Attention In VR Using Gaze Controlled Games. *Proceedings of* 2019 *ACM International Conference on Interactive Experiences for TV and Online Video* (*TVX*' 19), Salford, 5-7 June 2019, 167-173. https://doi.org/10.1145/3317697.3325118

[16] Wei, P., Liu, Y., Shu, T., Zheng, N. and Zhu, S.-C. (2018) Where and Why are They Looking? Jointly Inferring Human Attention and Intentions in Complex Tasks. *Proceedings of* 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Lake City, 18-23 June 2018, 6801-6809. https://doi.org/10.1109/CVPR.2018.00711

[17] Marin-Jimenez, M.J., Kalogeiton, V., Medina-Suarez, P. and Zisserman, A. (2019) LAEO-Net: Revisiting People Looking at Each Other in Videos. *Proceedings of the* 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 3472-3480. https://doi.org/10.1109/CVPR.2019.00359

[18] Yang, X., Xu, F., Wu, K., Xie, Z. and Sun, Y. (2021) Gaze-Aware Graph Convolutional Network for Social Relation Recognition. *IEEE Access*, **9**, 99398-99408. https://doi.org/10.1109/ACCESS.2021.3096553

[19] Zhuang, N., Ni, B., Xu, Y., Yang, X., Zhang, W., Li, Z. and Gao, W. (2019) Muggle: Multi-Stream Group Gaze Learning and Estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, **30**, 3637-3650. https://doi.org/10.1109/TCSVT.2019.2940479

[20] Recasens, A., Vondrick, C., Khosla, A. and Torralba, A. (2017) Following Gaze in Video. *Proceedings of* 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 1444-1452. https://doi.org/10.1109/ICCV.2017.160

[21] Brau, E., Guan, J., Jeffries, T. and Barnard, K. (2018) Multiple-Gaze Geometry: Inferring Novel 3D Locations from Gazes Observed in Monocular Video. In: Ferrari, V., Hebert, M., Sminchisescu, C. and Weiss, Y., Eds., *ECCV* 2018: *Computer Vision—ECCV* 2018, *Lecture Notes in Computer Science*, Vol. 11208, Springer, Cham, 612-630. https://doi.org/10.1007/978-3-030-01225-0_38

[22] Massé, B., Ba, S. and Horaud, R. (2017) Tracking Gaze and Visual Focus of Attention of People Involved in Social Interaction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **40**, 2711-2724. https://doi.org/10.1109/TPAMI.2017.2782819

[23] Long, M., Cao, Y., Wang, J. and Jordan, M. (2015) Learning Transferable Features with Deep Adaptation Networks. *Proceedings of the* 32*nd International Conference on Machine Learning*, Lille, 6-11 July 2015, 97-105.

[24] Xu, R., Li, G., Yang, J. and Lin, L. (2019) Larger Norm More Transferable: An Adaptive Feature Norm Approach for Unsupervised Domain Adaptation. *Proceedings of* 2019 *IEEE/CVF International Conference on Computer Vision Workshop*, Seoul, 27-28 October 2019, 1426-1435.

[25] Zen, G., Sangineto, E., Ricci, E. and Sebe, N. (2014) Unsupervised Domain Adaptation for Personalized Facial Emotion Recognition. *Proceedings of the* 16*th International Conference on Multimodal Interaction* (*ICMI*'14), Istanbul, 12-16 November 2014, 128-135. https://doi.org/10.1145/2663204.2663247

[26] Cui, S., Wang, S., Zhuo, J., Su, C., Huang, Q. and Tian, Q. (2020) Gradually Vanishing Bridge for Adversarial Domain Adaptation. *Proceedings of the* 16*th IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 12455-12464.

[27] Zhu, J.-Y., Park, T., Isola, P. and Efros, A.A. (2017) Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. *Proceedings of* 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 2242-2251.

[28] da Costa, T.V.G., Zara, G., Rota, P., Oliveira-Santos, T., Sebe, N., Murino, V. and Ricci, E. (2022) Dual-Head Contrastive Domain Adaptation for Video Action Recognition. *Proceedings of* 2020 *IEEE*/*CVF Winter Conference on Applications of Computer Vision* (*WACV*), Waikoloa, 3-8 January 2022, 1181-1190. https://doi.org/10.1109/WACV51458.2022.00229

[29] Wang, Q., Dai, D., Hoyer, L., Van Gool, L. and Fink, O. (2021) Domain Adaptive Semantic Segmentation with Self-Supervised Depth Estimation. *Proceedings of* 2021 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 8515-8525. https://doi.org/10.1109/ICCV48922.2021.00840

[30] Xu, J., Xiao, L. and López, A.M. (2019) Self-Supervised Domain Adaptation for Computer Vision Tasks. *IEEE Access*, **7**, 156694-156706. https://doi.org/10.1109/ACCESS.2019.2949697

[31] Kellnhofer, P., Recasens, A., Stent, S., Matusik, W. and Torralba, A. (2019) Gaze360: Physically Unconstrained Gaze Estimation in the Wild. *Proceedings of* 2019 *IEEE*/*CVF International Conference on Computer Vision*, Seoul, 27 October-2 November 2019, 6912-6921. https://doi.org/10.1109/ICCV.2019.00701

[32] Tzeng, E., Hoffman, J., Saenko, K. and Darrell, T. (2017) Adversarial Discriminative Domain Adaptation. *Proceedings of* 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 2962-2971. https://doi.org/10.1109/CVPR.2017.316

[33] Yu, Y., Liu, G. and Odobez, J.-M. (2019) Improving Few-Shot User-Specific Gaze Adaptation via Gaze Redirection Synthesis. *Proceedings of* 2019 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition*, Beach, 15-20 June 2019, 11937-11946. https://doi.org/10.1109/CVPR.2019.01221

[34] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M. and Lempitsky, V. (2016) Domain-Adversarial Training of Neural Networks. *The Journal of Machine Learning Research*, **17**, 2096-2030.

[35] Tomas, H., Reyes, M., Dionido, R., Ty, M., Mirando, J., Casimiro, J., Atienza, R. and Guinto, R. (2021) Goo: A Dataset for Gaze Object Prediction in Retail Environments. *Proceedings of* 2021 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition*, Nashville, 19-25 June 2021, 3125-3133. https://doi.org/10.1109/CVPRW53098.2021.00349

[36] Zhou, B., Lapedriza, A., Xiao, J., Torralba, A. and Oliva, A. (2014) Learning Deep Features for Scene Recognition Using Places Database. *Advances in Neural Information Processing Systems*, **27**, 487-495.

[37] Mora, K.A.F., Monay, F. and Odobez, J.-M. (2014) Eyediap: A Database for the Development and Evaluation of Gaze Estimation Algorithms from RGB and RGB-D Cameras. *Proceedings of* 2014 *Symposium on Eye Tracking Research and Applications* (*ETRA*' 14), Safety Harbor Florida, 26-28 March 2014, 255-258. https://doi.org/10.1145/2578153.2578190

[38] Judd, T., Ehinger, K., Durand, F. and Torralba, A. (2009) Learning to Predict Where Humans Look. *Proceedings of* 2009 *IEEE* 12*th International Conference on Computer Vision*, Kyoto, 29 September-02 October 2009, 2106-2113. https://doi.org/10.1109/ICCV.2009.5459462